XLSTAT 2023

https://www.xlstat.com

Introduction

XLSTAT est développé depuis plus de dix ans dans le but de rendre accessible au plus grand nombre un outil d'analyse de données et de statistique à la fois puissant, complet et convivial.

L'**accessibilité** vient de la compatibilité avec toutes les versions de Microsoft Excel aujourd'hui utilisées (Excel 2003 à Excel 2016), de l'interface disponible en 7 langues (allemand, anglais, chinois traditionnel, français, espagnol, italien, japonais, polonais et portugais) et de la mise à disposition sur le site www.xlstat.com d'une version d'évaluation utilisable 30 jours.

La **puissance** de XLSTAT vient à la fois du langage de programmation C++, et des algorithmes utilisés, qui sont le fruit des travaux de recherche de centaines de chercheurs statisticiens, mathématiciens ou informaticiens. Chaque développement d'une nouvelle fonctionnalité de XLSTAT est précédé d'une phase de recherche bibliographique approfondie, voire d'échanges avec les spécialistes des méthodes concernées.

La **complétude** de XLSTAT est le fruit d'une part de plus de dix ans de travail, et d'autre part d'échanges réguliers avec les utilisateurs, dont les idées et suggestions permettent de faire progresser le logiciel encore plus vite.

Enfin, la **convivialité** vient de l'interface, qui après quelques minutes de prise en main, rend facile et efficace l'utilisation de méthodes parfois très complexes qui requièrent dans d'autres logiciels des heures d'apprentissage.

L'architecture du logiciel a considérablement évolué au cours des 5 dernières années afin de prendre en compte les progrès d'Excel, et les problèmes de compatibilité entre les différentes plates-formes. Le logiciel s'appuie aujourd'hui sur le Visual Basic Application pour les interfaces et le C++ pour les calculs.

Comme toujours, les équipes d'Addinsoft et des distributeurs de XLSTAT se tiennent à votre disposition pour répondre à toute question, ou pour prendre en compte vos remarques et suggestions afin de continuer à améliorer le logiciel.

Installation

Licence

XLSTAT 2017 Contrat de Licence de l'Utilisateur Final

ADDINSOFT SARL ("ADDINSOFT") ACCEPTE DE VOUS CONCEDER LA LICENCE D'UTILISATION DE LA VERSION 2017 DE SON LOGICIEL XLSTAT ET DE LA DOCUMENTATION QUI L'ACCOMPAGNE (LE "LOGICIEL") A LA SEULE CONDITION QUE VOUS ACCEPTIEZ LES TERMES DE CE CONTRAT (LE « CONTRAT »). VEUILLEZ LIRE LES TERMES ATTENTIVEMENT. SI VOUS N'ACCEPTEZ PAS L'UN DES TERMES DE CE CONTRAT, ADDINSOFT REFUSE DE VOUS ACCORDER LA LICENCE D'UTILISATION DU LOGICIEL.

1. LICENCE. Par le présent contrat Addinsoft vous donne le droit non exclusif d'installer et d'utiliser le logiciel dans sa version électronique sur un seul ordinateur utilisable par un seul individu si vous utilisez la version de démonstration ou la version exempte de date limite d'utilisation. Si vous avez commandé une version multi-utilisateurs, le nombre d'utilisateurs dépend du nombre d'utilisateurs spécifié sur la facture qui a été transmise à vos services administratifs par Addinsoft.

2. RESTRICTIONS. Le Logiciel est la propriété intellectuelle d'Addinsoft et de ses fournisseurs. Tous les droits sur le logiciel qui ne font pas partie du contrat de licence sont entièrement réservés à Addinsoft. Vous n'avez pas le droit de décompiler, modifier ou utiliser les sources du logiciel pour toute utilisation non conforme aux lois en vigueur. Si une partie des sources devait vous apparaître par erreur vous devez impérativement en avertir Addinsoft. Toute tentative d'utilisation, de détournement ou de transfert de tout droit, devoir ou obligation mentionnés ci-dessous sera sans objet. Vous n'avez aucun droit de louer, revendre pour un profit quelconque le Logiciel. Vous n'avez aucun droit de reproduire ou distribuer le logiciel sans un accord préalable d'Addinsoft et hors du cadre prévu par l'article 1.

3. SUPPORT. Les utilisateurs enregistrés du Logiciel n'utilisant pas la version de démonstration ont le droit d'accéder au service après vente standard d'Addinsoft, les termes et les conditions de ce dernier pouvant être modifiés par Addinsoft à tout moment. Les utilisateurs de la version de démonstration peuvent contacter Addinsoft pour obtenir de l'aide sans toutefois avoir la garantie qu'il soit répondu à leurs demandes ou à leurs questions.

4. GARANTIE. Le Logiciel est livré "TEL QUEL" et Addinsoft rejette toute obligation de garantie concernant son Utilisation ou ses performances. Addinsoft et ses fournisseurs ne garantissent pas et ne peuvent pas garantir les performances ou les résultats que vous pouvez obtenir en

utilisant le logiciel. A l'exception de toute autre garantie, condition, représentation ou clause pour lesquelles les mêmes droits ne peuvent ou ne doivent pas être exclus ou limités par la loi applicable dans votre juridiction, Addinsoft et ses fournisseurs ne donnent aucune garantie, condition, représentation ou clause, expresse ou implicite, contractuelle, de droit commun, tirée de la coutume, ou des usages commerciaux ou autre, concernant d'autres sujets, y compris sans que ceci soit limitatif, concernant la non-violation des droits d'un tiers, la commercialisation, l'intégration du logiciel, sa qualité satisfaisante ou son adéquation à une fin spécifique.

5. LIMITATION DE RESPONSABILITÉ. En aucun cas Addinsoft et ses fournisseurs ne pourront être tenus pour responsable pour tous dommages, réclamations ou quelques coûts que ce soit ou pour tous dommages directs ou indirects, ou pour tout manque à gagner, pertes d'exploitation, pertes de bénéfices, et ce même si un représentant de Addinsoft a été informé de la possibilité de tels dommages, pertes, réclamations ou coûts. en aucun cas Addinsoft ou ses fournisseurs n'assument de responsabilité envers vous en cas de réclamation d'un tiers. Les limitations et restrictions ci-dessus s'appliquent dès lors qu'elles sont autorisées par la loi applicable dans votre juridiction. La responsabilité totale de Addinsoft ainsi que celle de ses fournisseurs dans le cadre de ce contrat ou en rapport avec ce dernier, est limitée à la somme versée pour l'acquisition du logiciel, s'il y a lieu. Aucune clause dans ce Contrat ne limite la responsabilité d'Addinsoft envers vous en cas de décès ou de préjudices corporels résultant d'une négligence avérée de la part d'Addinsoft. Addinsoft agit pour le compte de ses fournisseurs aux fins de réclamer, d'exclure et/ou de limiter les obligations, les garanties et les responsabilités stipulées dans ce Contrat, mais à aucun autre égard et dans aucun autre but.

6. CLAUSE RESOLUTOIRE. Ce Contrat est valable pour une durée maximum de 99 ans à moins qu'il n'y soit mis fin par l'une des deux parties. Vous pouvez mettre fin à ce contrat à tout moment en supprimant toutes les versions du logiciel sur l'ordinateur concerné. Ce contrat de licence sera résolu de fait si l'un des termes du présent contrat est violé. Après rupture du contrat vous serez obligé de supprimer toute copie du logiciel installé sur l'ordinateur concerné.

7. PARTIES CONTRACTANTES. Si ce logiciel est installé sur un ou des ordinateurs appartenant à une entreprise ou à toute autre personne morale, alors ce contrat est conclu entre Addinsoft et cette personne morale. La personne donnant son accord pour ce contrat certifie être habilitée à prendre l'engagement correspondant au présent contrat vis-à-vis d'Addinsoft.

8. INDEMNITES. Vous acceptez de vous engager à défendre Addinsoft contre toute plainte, réclamation, pertes, dommages, coûts et pertes, y compris les frais d'avocats, auxquels Addinsoft devrait faire face dans le cas de votre rupture du présent contrat.

9. GENERAL. Le logiciel est un « produit commercial ». Ce contrat est écrit dans l'esprit de la loi française et doit être interprété comme tel. En cas de contestation ou de litige, les juridictions attributives seront les tribunaux de Paris, France. Ce contrat est un contrat entre vous et

Addinsoft concernant le Logiciel et remplace tout autre accord préalable écrit et oral entre vous et Addinsoft au sujet du Logiciel.

PARIS, FRANCE, septembre 2017

Configuration minimale

XLSTAT fonctionne sous les systèmes d'exploitation suivants : Windows XP, Windows Vista, Windows 7, Windows 8.x, Windows 10, Mac OSX 10.6 à 10.12. Les plates-formes 32 et 64 bits sont supportées.

Pour fonctionner, XLSTAT a besoin que Microsoft Excel soit installé sur votre ordinateur. Les versions requises sur les systèmes Windows sont : Excel 2003 (11.0), Excel 2007 (12.0), Excel 2010 (14.0), Excel 2013 (15.0) et Excel 2016 (16.0). Sur le système Mac OSX Excel 2011 (SP1 ou supérieur) ou Excel 2016 sont requis.

Microsoft met régulièrement à votre disposition sur son site des patchs et des mises à jour des logiciels de la suite Office. Il est vivement recommandé d'installer ces mises à jour en raison des corrections parfois essentielles qu'elles comportent. Pour vérifier si votre version d'Excel est à jour, nous vous recommandons de vous rendre régulièrement sur :

https://docs.microsoft.com/fr-fr/officeupdates/

Installation

Pour installer XLSTAT vous devez :

- soit double-cliquer sur le fichier xlstat.exe (PC) ou xlstat.dmg (Mac) téléchargé depuis le site www.xlstat.com ou depuis le site de l'un de nos partenaires.

Si vos droits sont restreints sur l'ordinateur que vous utilisez, vous devez faire appel à un administrateur de la machine pour qu'il installe le logiciel. Une fois l'installation terminée, l'administrateur doit veiller à laisser un droit d'accès lecture/écriture aux éléments suivants :

- Dossier dans lequel se trouve Excel.exe
- Dossier dans lequel se trouve les fichiers utilisateur, (ex : C:\...\Application Data\Addinsoft\XLSTAT\)

Le répertoire pour les fichiers utilisateur pourra être changé ultérieurement par une personne ayant des droits d'administrateur sur l'ordinateur. Pour cela, il suffit d'utiliser l'option correspondante dans l'onglet « Avancées » de la boîte de dialogue des options XLSTAT.

Installation avancée

L'installation de XLSTAT sur un ordinateur hébergeant plusieurs comptes, sur un serveur, ou sur l'ensemble d'un parc informatique est simple à réaliser. Vous trouverez dans les sections ci-dessous des instructions pour vous guider.

Dans cette section :

Installation silencieuse avec InstallShield Script (Windows uniquement)

XLSTAT utilise un programme d'installation créé avec le logiciel InstallShield, basé uniquement sur le langage install script. Vous pouvez créer une installation silencieuse, comme avec n'importe quel logiciel s'appuyant sur InstallShield.

Pendant son installation, XLSTAT requière qu'une version de Microsoft Excel soit installée sur l'ordinateur. Excel est appelé une fois pendant l'installation par le programme d'installation afin d'ajouter le bouton de lancement rapide de XLSTAT. Il en est de même pendant la désinstallation, cette fois pour retirer le bouton de lancement rapide.

Utilisation d'un script InstallShield

Vous pouvez lancer le programme d'installation pour effectuer une installation silencieuse avec les options suivantes décrites dans l'aide d'InstallShield.

/uninst : cette option force une désinstallation de XLSTAT.

/s : l'installation sera effectuée sans montrer les boîtes de dialogue.

/f1 "fichier script" : ce paramètre indique quel fichier script doit être utilisé avec le chemin absolu et le nom complet de ce fichier.

/f2 "fichier log" : ce paramètre indique quel fichier log doit être utilisé avec le chemin absolu et le nom complet de ce fichier.

/r : ce paramètre active le mode enregistrement pour créer un fichier script.

/L : ce paramètre permet de choisir la langue du logiciel. Le tableau suivant donne les codes associés aux différentes langues proposées :

Option	Langue
/L1031	Allemand
/L1033	Anglais
/L1034	Espagnol
/L1036	Français
/L1040	Italien
/L1041	Japonais
/L1039	Portugais

/servername=XLSTATLICENSESERVER : ce paramètre indique le nom sur le réseau de la machine qui héberge le serveur de licence XLSTAT. Ce paramètre n'est utile que lors de l'installation silencieuse d'une machine cliente XLSTAT, dans le cadre d'un déploiement client/serveur. (remplacez XLSTATLICENSESERVER par le nom d'hôte de votre serveur de licence XLSTAT).

Une fois que l'installation de XLSTAT est terminée, deux fichiers script, l'un pour l'installation, l'autre pour la désinstallation, sont générés dans le répertoire silentinstall qui se trouve lui-même dans le répertoire d'installation de XLSTAT. Pour utiliser les scripts vous devez aussi disposer du fichier setup.exe qui se trouve dans le fichier xlstat.zip téléchargeable sur notre site.

Dans les exemples qui suivent, il est supposé, par confort, que les fichiers script et le fichier setup.exe se trouvent dans un même répertoire de travail C:\MyDir.

Installation silencieuse de XLSTAT

L'instruction pour lancer l'installation silencieuse peut être :

setup.exe /s /f1"C:\MyDir\setup.iss"

Dans ce cas le fichier de script setup.iss contient le texte suivant :

[InstallShield Silent]

Version=v7.00

File=Response File

[File Transfer]

OverwrittenReadOnly=NoToAll

[{68B36FA5-E276-4C03-A56C-EC25717E1668}-DlgOrder]

Dlg0={68B36FA5-E276-4C03-A56C-EC25717E1668}-SdWelcome-0

Count=9

Dlg1={68B36FA5-E276-4C03-A56C-EC25717E1668}-SdLicense2Rtf-0

Dlg2={68B36FA5-E276-4C03-A56C-EC25717E1668}-SetupType2-0

Dlg3={68B36FA5-E276-4C03-A56C-EC25717E1668}-SdAskDestPath2-0

Dlg4={68B36FA5-E276-4C03-A56C-EC25717E1668}-SdAskDestPath2-1

Dlg5={68B36FA5-E276-4C03-A56C-EC25717E1668}-SdComponentTree-0

Dlg6={68B36FA5-E276-4C03-A56C-EC25717E1668}-SdStartCopy2-0

Dlg7={68B36FA5-E276-4C03-A56C-EC25717E1668}-MessageBox-0

Dlg8={68B36FA5-E276-4C03-A56C-EC25717E1668}-SdFinish-0

[{68B36FA5-E276-4C03-A56C-EC25717E1668}-SdWelcome-0]

Result=1

[{68B36FA5-E276-4C03-A56C-EC25717E1668}-SdLicense2Rtf-0]

Result=1

[{68B36FA5-E276-4C03-A56C-EC25717E1668}-SetupType2-0]

Result=303

[{68B36FA5-E276-4C03-A56C-EC25717E1668}-SdAskDestPath2-0]

szDir=C:\Program Files\Addinsoft\XLSTAT

Result=1

[{68B36FA5-E276-4C03-A56C-EC25717E1668}-SdAskDestPath2-1]

szDir=C:\Mes documents\Addinsoft\

Result=1

[{68B36FA5-E276-4C03-A56C-EC25717E1668}-SdComponentTree-0]

szDir=C:\Program Files\Addinsoft\XLSTAT

Component-type=string

Component-count=4

Component-0=Program Files

Component-1=Help Files

Component-2=Icons & Menu

Component-3=SingleNode

Result=1

[{68B36FA5-E276-4C03-A56C-EC25717E1668}-SdStartCopy2-0]

Result=1

[{68B36FA5-E276-4C03-A56C-EC25717E1668}-MessageBox-0]

Result=1

[Application]

Name=XLSTAT 2017

Version=19.1.2810

Company=Addinsoft

Lang=040c

[{68B36FA5-E276-4C03-A56C-EC25717E1668}-SdFinish-0]

Result=1

bOpt1=0

bOpt2=0

Dans cet exemple, vous pouvez remplacer le répertoire d'installation "C:\Program Files\Addinsoft\XLSTAT" par un autre répertoire de votre choix. De même vous pouvez changer de répertoire utilisateur en remplaçant " C:\Mes documents\Addinsoft\" par un répertoire de votre choix.

Désinstallation silencieuse de XLSTAT

L'instruction pour lancer la désinstallation silencieuse peut être :

setup.exe /uninstall /s /f1"C:\MyDir\setupRemove.iss"

Dans ce cas le fichier de script setupRemove.iss contient le texte suivant :

[InstallShield Silent]

Version=v7.00

File=Response File

[File Transfer]

OverwrittenReadOnly=NoToAll

[{68B36FA5-E276-4C03-A56C-EC25717E1668}-DlgOrder]

Dlg0={68B36FA5-E276-4C03-A56C-EC25717E1668}-MessageBox-0

Count=2

Dlg1={68B36FA5-E276-4C03-A56C-EC25717E1668}-SdFinish-0

[{68B36FA5-E276-4C03-A56C-EC25717E1668}-MessageBox-0]

Result=6

[{68B36FA5-E276-4C03-A56C-EC25717E1668}-SdFinish-0]

Result=1

bOpt1=0

bOpt2=0

Installation silencieuse de XLSTAT Server pour les licences concurrentes serveur

L'instruction pour lancer l'installation silencieuse peut être :

setup.exe /s /f1"C:\MyDir\setup.iss"

Dans ce cas le fichier de script setup.iss contient le texte suivant :

[InstallShield Silent]

Version=v7.00

File=Response File

[File Transfer]

OverwrittenReadOnly=NoToAll

[{68B36FA5-E276-4C03-A56C-EC25717E1668}-DlgOrder]

Dlg0={68B36FA5-E276-4C03-A56C-EC25717E1668}-SdWelcome-0

Count=8

Dlg1={68B36FA5-E276-4C03-A56C-EC25717E1668}-SdLicense2Rtf-0

Dlg2={68B36FA5-E276-4C03-A56C-EC25717E1668}-SetupType2-0

Dlg3={68B36FA5-E276-4C03-A56C-EC25717E1668}-SdAskDestPath2-0

Dlg4={68B36FA5-E276-4C03-A56C-EC25717E1668}-SdAskDestPath2-1

Dlg5={68B36FA5-E276-4C03-A56C-EC25717E1668}-SdComponentTree-0

Dlg6={68B36FA5-E276-4C03-A56C-EC25717E1668}-SdStartCopy2-0

Dlg7={68B36FA5-E276-4C03-A56C-EC25717E1668}-SdFinish-0

[{68B36FA5-E276-4C03-A56C-EC25717E1668}-SdWelcome-0]

Result=1

[{68B36FA5-E276-4C03-A56C-EC25717E1668}-SdLicense2Rtf-0]

Result=1

[{68B36FA5-E276-4C03-A56C-EC25717E1668}-SetupType2-0]

Result=303

[{68B36FA5-E276-4C03-A56C-EC25717E1668}-SdAskDestPath2-0]

szDir=C:\Program Files\Addinsoft\XLSTAT

Result=1

[{68B36FA5-E276-4C03-A56C-EC25717E1668}-SdAskDestPath2-1]

szDir= C:\Mes documents\Addinsoft\

Result=1

[{68B36FA5-E276-4C03-A56C-EC25717E1668}-SdComponentTree-0]

szDir=C:\Program Files\Addinsoft\XLSTAT

Component-type=string

Component-count=5

Component-0=Program Files

Component-1=Help Files

Component-2=Icons & Menu

Component-3=Server setup

Component-4=SingleNode

Result=1

[{68B36FA5-E276-4C03-A56C-EC25717E1668}-SdStartCopy2-0]

Result=1

[Application]

Name=XLSTAT 2017

Version=19.1.2810

Company=Addinsoft

Lang=040c

[{68B36FA5-E276-4C03-A56C-EC25717E1668}-SdFinish-0]

Result=1

bOpt1=0

bOpt2=0

Dans cet exemple, vous pouvez remplacer le répertoire d'installation "C:\Program Files\Addinsoft\XLSTAT" par un autre répertoire de votre choix. De même vous pouvez changer de répertoire utilisateur en remplaçant "C:\Mes documents\Addinsoft\" par un répertoire de votre choix.

Installation silencieuse du client serveur XLSTAT Client dans le cas d'une licence concurrente serveur

L'instruction pour lancer l'installation silencieuse peut être :

setup.exe /s /f1"C:\MyDir\setup.iss"

Dans ce cas le fichier de script setup.iss contient le texte suivant :

[InstallShield Silent]

Version=v7.00

File=Response File

[File Transfer]

OverwrittenReadOnly=NoToAll

[{68B36FA5-E276-4C03-A56C-EC25717E1668}-DlgOrder]

Dlg0={68B36FA5-E276-4C03-A56C-EC25717E1668}-SdWelcome-0

Count=9

Dlg1={68B36FA5-E276-4C03-A56C-EC25717E1668}-SdLicense2Rtf-0

Dlg2={68B36FA5-E276-4C03-A56C-EC25717E1668}-SetupType2-0

Dlg3={68B36FA5-E276-4C03-A56C-EC25717E1668}-SdAskDestPath2-0

Dlg4={68B36FA5-E276-4C03-A56C-EC25717E1668}-SdAskDestPath2-1

Dlg5={68B36FA5-E276-4C03-A56C-EC25717E1668}-SdComponentTree-0

Dlg6={68B36FA5-E276-4C03-A56C-EC25717E1668}-AskText-0

Dlg7={68B36FA5-E276-4C03-A56C-EC25717E1668}-SdStartCopy2-0

Dlg8={68B36FA5-E276-4C03-A56C-EC25717E1668}-SdFinish-0

[{68B36FA5-E276-4C03-A56C-EC25717E1668}-SdWelcome-0]

Result=1

[{68B36FA5-E276-4C03-A56C-EC25717E1668}-SdLicense2Rtf-0]

Result=1

[{68B36FA5-E276-4C03-A56C-EC25717E1668}-SetupType2-0]

Result=303

[{68B36FA5-E276-4C03-A56C-EC25717E1668}-SdAskDestPath2-0]

szDir=C:\Program Files\Addinsoft\XLSTAT

Result=1

[{68B36FA5-E276-4C03-A56C-EC25717E1668}-SdAskDestPath2-1]

szDir= C:\Mes documents\Addinsoft\

Result=1

[{68B36FA5-E276-4C03-A56C-EC25717E1668}-SdComponentTree-0]

szDir=C:\Program Files\Addinsoft\XLSTAT

Component-type=string

Component-count=5

Component-0=Program Files

Component-1=Help Files

Component-2=Icons & Menu

Component-3=Client setup

Component-4=SingleNode

Result=1

[{68B36FA5-E276-4C03-A56C-EC25717E1668}-AskText-0]

szText=XLSTATLICENSESERVER

Result=1

[{68B36FA5-E276-4C03-A56C-EC25717E1668}-SdStartCopy2-0]

Result=1

[Application]

Name=XLSTAT 2017

Version=19.1.2810

Company=Addinsoft

Lang=040c

[{68B36FA5-E276-4C03-A56C-EC25717E1668}-SdFinish-0]

Result=1

bOpt1=0

bOpt2=0

Dans cet exemple, vous pouvez remplacer le répertoire d'installation "C:\Program Files\Addinsoft\XLSTAT" par un autre répertoire de votre choix. De même vous pouvez changer de répertoire utilisateur en remplaçant "C:\Mes documents\Addinsoft\" par un répertoire de votre choix.

Vous devez renseigner le nom du serveur de licence XLSTAT tel qu'il apparaît sur votre réseau, en remplaçant "XLSTATLICENSESERVER" par le nom de la machine sur laquelle XLSTAT Server est installé.

Création d'un fichier script personnalisé

Pour modifier plus en profondeur l'installation, vous pouvez enregistrer une installation manuelle afin de créer un fichier script réutilisable par la suite. Pour cela, utilisez l'option /r. Voici un exemple d'appel générant un fichier setup.iss :

setup.exe /r /f1"C:\MyDir\setup.iss"

Choix de la langue

Si XLSTAT est installé pour la première fois et si l'installation est réalisée classiquement avec l'interface utilisateur, la langue par défaut sera celle choisie par la personne installant le logiciel.

L'utilisateur pourra néanmoins à tout moment changer de langue en utilisant la boîte de dialogue des Options de XLSTAT. Une démonstration est disponible sur

http://www.xlstat.com/demo-langf.htm

Si XLSTAT est installé pour la première fois sur l'ordinateur, et que l'installation se fait en mode silencieux, alors l'anglais est la langue par défaut. Il existe deux possibilités de changer la langue par défaut de XLSTAT avant l'installation :

- /L: utilisez cette option dans l'instruction de lancement de l'installation silencieuse pour choisir la langue de XLSTAT. Le tableau suivant donne les codes associés aux différentes langues proposées :

Option	Langue
/L1031	Allemand
/L1033	Anglais
/L2052	Chinois simplifié
/L1028	Chinois traditionnel
/L1034	Espagnol
/L1036	Français
/L1040	Italien
/L1041	Japonais
/L1045	Polonais
/L2070	Portugais

- **Entrée de la base de registre** : une fois que l'installation de XLSTAT est terminé, si XLSTAT n'a pas encore été lance, vous pouvez modifier la langue au niveau de la base de registre en changeant la valeur de la clef HKEY_LOCAL_MACHINE\SOFTWARE\XLSTAT+\General\Language vers l'une des valeurs ci-dessous :

Code	Language
DE	Allemand
US	Anglais
CS	Chinois simplifié
CN	Chinois traditionnel
ES	Espagnol
FR	Français
IT	Italien
JP	Japonais
PL	Polonais
PT	Portugais

Si XLSTAT a déjà été installé sur l'ordinateur, le choix de la langue au moment de l'installation avec l'option /L ou en modifiant l'entrée dans la base de registre, sera sans effet. Chaque utilisateur retrouvera le choix qu'il a précédemment fait dans XLSTAT (il peut s'agir du choix fait lors de la première installation si aucune modification n'a été effectuée depuis). L'utilisateur peut néanmoins à tout moment changer de langue en utilisant la boîte de dialogue des Options de XLSTAT.

Choix du répertoire utilisateur

XLSTAT permet à l'utilisateur d'enregistrer les sélections de données et les différentes options des boîtes de dialogue pour les réutiliser par la suite. La façon dont cela est mémorisé est paramétrable dans la boîte des Options de XLSTAT. Le répertoire utilisateur peut être défini ou modifié au niveau des Options de XLSTAT (onglet Avancées), mais il est plus pratique de le définir une fois pour toute au moment de l'installation.

Installation standard de XLSTAT

Le répertoire utilisateur de XLSTAT est choisi par InstallShield pendant l'installation comme étant :

%USERPROFILE%\Application data\ADDINSOFT\XLSTAT

où la variable d'environnement %USERPROFILE% est remplacée par sa valeur au moment de l'installation.

Il est possible de modifier ce répertoire pendant l'installation en modifiant cette valeur, ou après l'installation soit en utilisant la boîte des Options de XLSTAT (onglet Avancées), soit en modifiant la valeur de la clef

HKEY_CURRENT_USER\Software\XLSTAT+\DATA\UserPath

Cette clef a priorité sur le choix effectué au niveau de la boîte de dialogue des Options et peut contenir des variables d'environnement.

Environnement multi-utilisateurs

Il existe différent d'environnements multi-utilisateurs. Par exemple, cela peut être une installation sur un serveur Windows Terminal Server ou Citrix Metaframe Server. Cela peut aussi être le cas d'un ensemble d'ordinateurs où l'environnement est installé à partir d'une image répliquée sur tous les ordinateurs et où les utilisateurs ont accès sur chaque machine à XLSTAT. Dans de tels cas, veuillez prendre en compte la recommandation suivante : le répertoire utilisateur doit pointer vers un répertoire personnel sur lequel l'utilisateur a des droits de lecture et écriture. Il y a deux moyens de permettre cela :

- Utiliser un répertoire virtuel
- Utiliser des variables d'environnement

Répertoire virtuel

Si vous utilisez un répertoire utilisateur virtuel, celui-ci doit exister préalablement. Ce répertoire doit avoir le même nom pour tous les utilisateurs mais pointer sur des répertoires physiques différents. Un répertoire utilisateur virtuel est souvent désigné par un chemin tel que U:\ ou X:\. Au moment où l'utilisateur se logue, le répertoire virtuel est souvent monté automatiquement par un script. Les utilisateurs ont en principe des droits de lecture et écriture sur ce répertoire. Pour XLSTAT cela est suffisant.

Si le répertoire virtuel est par exemple **U:**, alors vous pouvez choisir comme répertoire utilisateur pour XLSTAT le répertoire suivant qui respecte les conventions Microsoft :

U:\Application Data\ADDINSOFT\XLSTAT

Ce répertoire doit exister pour tous les utilisateurs potentiels de XLSTAT, avant que XLSTAT ne soit utilisé par ces utilisateurs. Si cela n'est pas le cas, un message d'erreur informe l'utilisateur que ce répertoire est absent, et propose à l'utilisateur de choisir un autre répertoire.

Variables d'environnement

Si vous choisissez de définir le répertoire utilisateur en utilisant des variables d'environnement, la valeur de la variable d'environnement doit permettre d'identifier un répertoire utilisateur existant et sur lequel l'utilisateur a des droits de lecture et écriture.

Par exemple la variable d'environnement **%USERPROFILE%** peut être utilisée pour définir le répertoire utilisateur suivant, qui respecte les conventions Microsoft :

%USERPROFILE%\Application Data\ADDINSOFT\XLSTAT

L'utilisation de variables d'environnement dans l'interface d'installation d'InstallShield n'est pas autorisée. Vous pouvez en revanche utiliser des variables d'environnement dans un script ou dans les entrées de la base de registre.

Installation sur un serveur et création d'une image d'installation

L'installation sur un serveur et la création d'une image doit pouvoir être réalisé sans problème. Veuillez noter que Microsoft Excel doit avoir été préalablement installé sur l'ordinateur avec toutes les options pour le VBA, les macros, Microsoft Forms et les filtres graphiques. Pendant une installation sur un serveur Windows Terminal Server, Microsoft Excel version 2003 ou suivante est recommandé.

Pendant l'installation de XLSTAT, il est nécessaire que le répertoire dans lequel se trouve le fichier Excel.exe soit en accès libre pour la lecture et l'écriture.

Si vous avez des questions au sujet de l'installation sur un serveur, n'hésitez pas à contacter le service support d'Addinsoft.

Bibliographie

InstallShield 2008 Help Library. Setup.exe and Update.exe Command-Line Parameters, http://helpnet.acresso.com/robo/projects/installshield14helplib/IHelpSetup_EXECmdLine.htm, Macrovision.

Utiliser XLSTAT

L'esprit de XLSTAT

XLSTAT est un logiciel dont l'interface s'appuie entièrement sur Microsoft Excel, tant pour la récupération des données que pour la restitution des résultats. Les calculs sont en revanche totalement indépendants de Microsoft Excel et ont été développés avec le langage de programmation C++.

Afin de vous garantir une qualité irréprochable des résultats proposés, le logiciel XLSTAT a fait l'objet de tests intensifs, et a été validé par des spécialistes des méthodes utilisées.

Dans un souci d'amélioration permanente des logiciels qu'elle propose, la société Addinsoft est à l'écoute des remarques et suggestions que vous voudriez lui transmettre. Pour contacter Addinsoft, vous pouvez écrire à support@xlstat.com.

Sélection des données

Comme pour l'ensemble des modules XLSTAT, la sélection des données se fait directement sur la feuille Excel, de préférence avec la souris. Les logiciels de statistique affichent classiquement des listes de variables à sélectionner ou non pour la méthode employée ou non. L'approche de XLSTAT est complètement différente puisque vous choisissez les données directement sur une ou plusieurs feuilles Excel.

Deux modes de sélection sont à votre disposition, sachant que pour chaque variable ou groupe de variables (par exemple d'une part la variable dépendante, d'autre part les variables quantitatives explicatives) vous pouvez opter pour l'un des modes. Les deux modes sont :

- **Sélection par plage** : vous sélectionnez avec la souris l'ensemble des cellules de la feuille Excel correspondant aux variables ou au tableau de données, après avoir cliqué dans la zone correspondante de la boîte de dialogue.

- **Sélection par colonnes** : ce mode de sélection ne peut être utilisé que si votre tableau de données commence sur la première ligne de la feuille Excel. Après avoir cliqué dans la zone de la boîte de dialogue correspondant à la sélection que vous voulez faire, vous devez cliquer sur le nom de la première colonne correspondant à votre tableau (A, B, C, …), puis sélectionner les autres colonnes en laissant le bouton droit de la souris enfoncé.

- **Sélection par lignes** : ce mode de sélection ne peut être utilisé que si votre tableau de données commence sur la première colonne de la feuille Excel (colonne A). Après avoir cliqué dans la zone de la boîte de dialogue correspondant à la sélection que vous voulez faire, vous devez cliquer sur le nom de la première ligne correspondant à votre tableau (1, 2, 3, …), puis sélectionner les autres lignes en laissant le bouton droit de la souris enfoncé.

Remarques :

- Les sélections multiples sont possibles. Par exemple, si vos variables vont de la colonne B à la colonne G, mais que vous ne souhaitez pas inclure la colonne E dans l'analyse, vous pouvez sélectionner les colonnes B à D avec la souris, puis cliquer sur la touche Ctrl puis sélectionner les colonnes F à G en laissant la touche Ctrl enfoncée. Vous pouvez aussi sélectionner les colonnes B à G, puis cliquer sur la touche Ctrl puis sélectionner la colonne E.

- Vous ne pouvez pas mélanger les modes à l'intérieur d'une sélection. En revanche, vous pouvez utiliser différents modes à l'intérieur d'une même boîte de dialogue.

- Si votre tableau comprend sur la première ligne le libellé des variables, vous devez veiller à ce que l'option « Libellés des variables », « Libellés des colonnes » ou « Libellés présents » soit activée.

- Vous pouvez utiliser les raccourcis clavier pour sélectionner les données très rapidement et sans la souris. Cela n'est toutefois possible que si vous avez installé les derniers correctifs pour Excel. Les raccourcis les plus utiles sont les suivants :

- **Ctrl A** : sélectionne toutes les cellules de la feuille

- **Ctrl Space** : sélectionne toute la colonne correspondant aux cellules déjà sélectionnées
- **Shift Space** : sélectionne toute la ligne correspondant aux cellules déjà sélectionnées
- Quand la sélection active correspond à une cellule ou à un groupe de cellules :
- **Shift Bas** : sélectionne sur une ligne vers le bas les cellules adjacentes aux cellules déjà sélectionnées
- **Shift Haut** : sélectionne sur une ligne vers le haut les cellules adjacentes aux cellules déjà sélectionnées
- **Shift Gauche** : sélectionne sur une colonne vers la gauche les cellules adjacentes aux cellules déjà sélectionnées
- **Shift Droite** : sélectionne sur une colonne vers la droite les cellules adjacentes aux cellules déjà sélectionnées
- **Ctrl Shift Bas** : sélectionne vers le bas toutes les cellules non vides adjacentes aux cellules déjà sélectionnées
- **Ctrl Shift Haut** : sélectionne vers le haut toutes les cellules non vides adjacentes aux cellules déjà sélectionnées
- **Ctrl Shift Gauche** : sélectionne vers la gauche toutes les cellules non vides adjacentes aux cellules déjà sélectionnées
- **Ctrl Shift Droite** : sélectionne vers la droite toutes les cellules non vides adjacentes aux cellules déjà sélectionnées
- Quand la sélection active correspond à une ou plusieurs colonnes :
- **Shift Gauche** : sélectionne la colonne à gauche des colonnes déjà sélectionnées
- **Shift Droite** : sélectionne la colonne à droite des colonnes déjà sélectionnées
- **Ctrl Shift Gauche** : sélectionne vers la gauche toutes les colonnes non vides adjacentes aux colonnes déjà sélectionnées
- **Ctrl Shift Droite** : sélectionne vers la droite toutes les colonnes non vides adjacentes aux colonnes déjà sélectionnées
- Quand la sélection active correspond à une ou plusieurs lignes :
- **Shift Bas** : sélectionne la ligne à gauche des lignes déjà sélectionnées
- **Shift Haut** : sélectionne la ligne à droite des lignes déjà sélectionnées
- **Ctrl Shift Bas** : sélectionne vers le bas toutes les lignes non vides adjacentes aux lignes déjà sélectionnées
- **Ctrl Shift Haut** : sélectionne vers le haut toutes les lignes non vides adjacentes aux lignes déjà sélectionnées

Voir aussi :

http://www.xlstat.com/demo-selectf.htm

Messages

XLSTAT vous propose un système innovant et performant pour la gestion des messages d'information et des messages d'erreur. La boîte ci-dessous présente un exemple de message qui se produit dans le cas où un champ de sélection de données est vide (en l'occurrence les variables dépendantes), alors qu'une sélection est attendue.

La zone en rouge (ou en bleu en fonction du contexte) vous indique quel champ de la boîte de dialogue est concerné. Si vous cliquez sur Retour, le champ concerné est automatiquement activé.

Le message est en principe explicite et devrait vous aider à résoudre le problème rapidement. Si ce n'était toutefois pas le cas, en cliquant sur l'hyperlien « http://www.xlstat.com » vous pouvez vous connecter sur le site de XLSTAT et accéder au tutoriel le plus pertinent. Vous pouvez aussi transmettre le message à XLSTAT, soit en cliquant sur l'adresse « support@xlstat.com » qui apparaît parfois sous l'hyperlien, soit en copiant le message en faisant « Alt – Shift – Impr Ecran » puis en collant le message dans un email (en cliquant Ctrl C par exemple).

Options générales

Options

XLSTAT offre un nombre important d'options afin de vous permettre une utilisation personnalisée et optimale du logiciel.

Pour afficher la boîte de dialogue des options de XLSTAT, cliquez sur la commande « Options » du menu XLSTAT ou cliquez sur le bouton  de la barre d'outils XLSTAT.

: cliquez sur ce bouton pour enregistrer les modifications.

: cliquez sur ce bouton pour fermer la boîte de dialogue. Si vous n'avez pas préalablement enregistré vos modifications, elles ne seront pas prises en compte.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.

Onglet **Générales** :

Langue : utilisez cette option pour modifier la langue de l'interface de XLSTAT.

Focus : cette option n'est visible que si vous utilisez la version d'essai ou la version Premium. Dans ce cas, vous pouvez utiliser cette option pour afficher un bouton donnant accès plus rapide à une séries de fonctions correspondant à l'un des champs couverts par les solutions "Applied" de XLSTAT : Forecasting, LifeScience, Marketing, Quality, Sensory.

Entrées des boîtes de dialogue :

- **Mémoriser pendant une s ession** : activez cette option si vous souhaitez que XLSTAT mémorise le temps d'une session (ouverture / fermeture de XLSTAT) les différentes entrées des boîtes de dialogue.
- **Y compris pour les sélections de données** : activez cette option si vous souhaitez que XLSTAT conserve pendant une session les sélections de données.
- **Mémoriser d'une s ession à l'autre** : activez cette option si vous souhaitez que XLSTAT mémorise les différentes entrées des boîtes de dialogue d'une session à l'autre.
- **Y compris pour les sélections de données** : activez cette option si vous souhaitez que XLSTAT conserve aussi d'une session à l'autre les sélections de données. Cette option est particulièrement utile si vous travaillez souvent sur des feuilles Excel qui ont le même nom et une structure de données identiques.

Demander la confirmation des sélections : activez cette option si vous souhaitez que XLSTAT vous demande de confirmer les sélections de données après que vous avez cliqué sur le bouton OK des boîtes de dialogue. Si vous activez cette option, vous aurez la possibilité de vérifier le nombre de lignes et de colonnes sélectionnées pour l'ensemble des sélections actives.

Me prévenir avant que la licence ou l'accès aux mises à jour n'expire : activez cette option si vous souhaitez être prévenu par XLSTAT avant que votre licence n'expire (licence annuelle) ou avant que votre accès gratuit aux mises à jour n'expire (licence perpétuelle).

Montrer seulement les fonctions actives dans les menus et les barres d'outils : Activez cette option si vous souhaitez que seules les fonctions actives correspondant à des modules auxquels la licence donne accès soient affichées dans le menu XLSTAT et les barres d'outils.

Afficher les messages d'information : activez cette option si vous souhaitez recevoir les messages d'information au démarrage de XLSTAT.

Afficher les barres d'outils XLSTAT dans l'onglet Compléments : activez cette option pour afficher les barres d'outils de XLSTAT dans l'onglet Compléments d'Excel (Excel 2007 et suivantes uniquement).

Onglet **Données** :

Filtres Excel : utilisez les options suivantes pour définir la façon dont XLSTAT doit gérer les filtres, lorsqu'ils existent, qui ont été appliqués à votre feuille de calcul.

- **Demander à l'utilisateur** : activez cette option si vous souhaitez que XLSTAT vous consulte à chaque fois qu'un filtre est trouvé. Vous aurez alors le choix entre utiliser les filtres tels quels ou les ignorer.
- **Appliquer les filtres tels qu'ils sont dans la feuille de calcul** : appliquer les filtres tels qu'ils sont dans la feuille de calcul.
- **Ignorer les filtres** : Activez cette option si vous souhaitez que XLSTAT ignore les filtres et utilise l'ensemble de données.

Données manquantes:

Considérer les cellules vides comme des données manquantes : cette option est active par défaut et ne peut être désactivée. XLSTAT considère systématiquement qu'une cellule vide dans une sélection correspond à une donnée manquante.

Considérer aussi les valeurs suivantes comme des données manquantes : si vous activez cette option, les valeurs indiquées dans la liste en dessous de l'option seront aussi considérées comme des données manquantes, que ce soit pour des données numériques ou des données nominales.

Considérer toute donnée textuelle comme une donnée manquante : cette option ne s'applique qu'aux sélections de données numériques. Quelle que soit la donnée textuelle rencontrée, elle sera considérée comme une donnée manquante. Si vous activez cette option soyez sûr que des données n'ont pas été converties par mégarde d'un format numérique en un

format texte : vous risqueriez d'ignorer des observations alors qu'une rectification vous permettrait de les inclure dans les calculs.

Onglet **Sorties**:

Style : choisissez le style que vous préférez pour les sorties parmi « Classique » qui correspond au format historique de XLSTAT, « Moderne » qui correspond à un autre jeu de couleurs et « Scientifique » qui n'utilise que du noir du blanc et des gris.

Position des nouvelles feuilles : si vous choisissez l'option de sortie « Feuille » dans les boîtes de dialogue des fonctions XLSTAT, utilisez cette option pour modifier la position des feuilles de résultats dans le classeur Excel.

Nombre de décimales : choisissez le nombre de décimales à afficher pour les résultats numériques. Notez que vous avez toujours la possibilité de voir par la suite un nombre de décimales inférieur ou supérieur en utilisant les options de formatage d'Excel.

p-value minimale : entrez la valeur p-value minimale en-dessous de laquelle la p-value est remplacée par « < p » où p est la p-value minimale.

Colorer les onglets : activez cette option et choisissez une couleur si vous souhaitez que les onglets générés par XLSTAT dans votre classeur aient une couleur particulière.

Afficher les titres en gras : activez cette option pour que XLSTAT affiche les titres des tableaux de résultats en gras.

Lignes vides après les titres : choisissez le nombre de lignes à laisser vides après un titre. Le nombre de lignes laissées vides après un tableau ou un graphique correspond à ce nombre +1.

Afficher l'en-tête des tableaux en gras : activez cette option pour que XLSTAT affiche en-têtes des tableaux de résultats en gras.

Afficher la liste des résultats dans l'en-tête du rapport : activez cette option pour que XLSTAT affiche la liste des tableaux et graphiques de résultats dans l'en-tête du rapport.

Afficher le nom du projet dans l'en-tête du rapport : activez cette option pour que XLSTAT affiche le nom de votre projet dans l'en-tête du rapport, puis entrez le nom de votre projet dans le champ correspondant.

Elargir la première colonne du rapport par un facteur de X : activez cette option pour élargir automatiquement la première colonne du rapport de XLSTAT d'un facteur X. La valeur par défaut est 1, et correspond à laisser la largeur de la colonne inchangée.

Onglet **Graphiques** :

Afficher les graphiques sur des feuilles séparées : activez cette option pour que les graphiques soient affichés sur des feuilles graphiques séparées. Remarque : lorsque des graphiques sont affichés sur une feuille Excel standard, vous pouvez les convertir en feuille graphique séparée en les sélectionnant, puis en faisant un clic droit avec votre souris, puis en cliquant sur « Emplacement », puis en choisissant « sur une nouvelle feuille ».

Taille des graphiques :

- **Automatique** : choisissez cette option si vous souhaitez que XLSTAT détermine automatiquement la taille des graphiques en utilisant comme point de départ la hauteur et la largeur définies ci-dessous.
- **Définie par l'utilisateur** : activez cette option si vous souhaitez que XLSTAT affiche des graphiques dont la taille est exactement définie par les valeurs ci-dessous :
- **Largeur** : entrez la valeur en points de la largeur des graphiques ;
- **Hauteur** : entrez la valeur en points de la hauteur des graphiques.

Afficher des graphiques orthonormés : activez cette option pour que les graphiques issus d'analyses factorielles soient orthonormés. Cela permet d'avoir automatiquement des échelles identiques pour les abscisses et les ordonnées, et d'éviter des interprétations erronées du fait d'effets de dilatation artificiels.

Onglet **Avancées** :

Nombres aléatoires :

Fixer la graine à : activez cette option si vous voulez vous assurer que les résultats mettant en jeu des calculs sur des nombres aléatoires donnent toujours le même résultat. Entrez alors la valeur de la graine (le point de départ de génération des nombres aléatoires).

Nombre maximum de processeurs : XLSTAT est un logiciel capable de paralléliser les calculs. Vous pouvez ici choisir le nombre de processeurs auxquels XLSTAT peut accéder.

Utiliser les GPU NVIDIA : GPU signifie processeurs graphiques (ou Graphical Processing Unit en anglais). Ces cartes sont très répandues dans les appareils informatiques d'aujourd'hui où elles permettent de réaliser très rapidement les calculs de rendu et d'affichage graphique haute définition pour de nombreuses applications. Une seconde possibilité consiste à les utiliser comme calculateurs génériques sur processeur graphique (ou GPGPU pour general- purpose processing units en anglais) au sein d'algorithmes complexes pour faire ce en quoi elles excellent : réaliser des volumes de calculs importants à une vitesse proprement hallucinante.

XLSAT a choisi NVIDIA, le fabriquant des GPUS les plus répandus et les plus puissants, pour implémenter sur GPU un nombre croissant de ses algorithme afin d'offrir de meilleures performances et des économies d'énergies à ses utilisateurs. Les méthodes ayant été implémentées sur GPU sont reconnaissables par la ligne "accéléré par GPU" dans leur description de l'aide d'XLSTAT.

Si votre appareil est équipé de processeurs graphiques NVIDIA et si vous utilisez la version 64 bits d'Excel, vous pouvez cocher cette option pour activer l'accélération GPU sur les algorithmes possédant une implémentation GPU. Vous allez alors observer une accélération significative de vos méthodes habituelles.

Afficher les boutons avancés dans les boîtes de dialogue : Activez cette option si vous souhaitez afficher les boutons permettant d'enregistrer et charger les paramètres des boîtes de dialogue pour de futures utilisations, ainsi que le bouton permettant de générer le code VBA pour automatiser le lancement de procédures XLSTAT.

Chemin pour les fichiers utilisateurs : vous pouvez modifier le répertoire dans lequel doivent être enregistrés les fichiers utilisateurs en cliquant sur le bouton […] qui vous permettra de choisir le répertoire. Les fichiers utilisateurs comprennent les options définies dans cette boîte de dialogue et les options des boîtes de dialogues des différents outils. Le répertoire dans lequel sont enregistrés ces fichiers doit être accessible en lecture/écriture.

Filières

La fonctionnalité des filières permet de combiner et exécuter plusieurs analyses à la suite. Chacune peut utiliser les résultats des analyses précédentes comme données d'entrée ainsi que des données provenant des classeurs Excel ouverts. L'enchainement des analyses de données et statistiques est ainsi fluidifié et est présenté de manière plus simple. Cela permet d'avoir une vision d'ensemble des analyses tout en gardant la possibilité de rejouer chaque analyse de manière indépendante. Des fonctionnalités d'export/import des filières permettent de partager simplement les configurations.

Dans cette section :

Description

Une filière est une succession d'analyses représentées par des blocs connectés entre eux. Visuellement, cela ressemble à un arbre qui évolue de gauche à droite. Les premiers blocs (à gauche) correspondent aux données d'entrée et tous les suivants sont les analyses qui se chaînent les unes après les autres, dans l'ordre représenté par l'arbre. Les derniers blocs sont le résultat final de la filière. Pour chaque bloc analyse, les données d'entrées peuvent être sélectionnées parmi les données de sortie de tous les blocs précédents. Il suffit qu'ils apparti140ennent à une même branche.

Données d'entrée : 

- Vous pouvez sélectionner celles-ci, via le bloc relatif aux données d'entrée, parmi toutes les plages de données des classeurs Excel ouverts. Cette sélection peut être manuelle ou automatique. Un outil vous permet de détecter automatiquement les données des classeurs ouverts et de les réutiliser pour les analyses à venir.
- Aucun bloc ne peut précéder un bloc relatif aux données d'entrée.
- Si des données sont filtrées alors cela n'empêche pas leur sélection. Selon vos paramétrages dans les options XLSTAT, les filtres seront appliqués ou non.
- Vous pouvez choisir d'avoir un nombre de lignes fixe ou non. Cela permet de pouvoir rester sur une sélection de données précise ou de laisser l'outil ajouter/enlever des lignes selon vos actions sur les données d'entrée.
- La même chose est possible pour les colonnes. Cependant, si vous supprimez des colonnes qui sont utilisées dans la suite de votre filière alors les blocs concernés et ceux impactés seront réinitialisés avec suppression de la feuille résultat si elle existe.

Voir la patie Création d'une nouvelle filière pour avoir plus de détails sur ces différents points.

Différentes analyses utilisables

Vingt-cinq analyses sont actuellement disponibles pour créer une filière. Elles peuvent toutes se succéder. Cependant, les Statistiques descriptives, les Histogrammes et les Nuages de points ne peuvent avoir de suite, ce sont nécessairement des blocs finaux.

- **Préparation des données** :
  -  : Données manquantes
  -  : Gestion des données
  -  : Anonymisation des données

-  : Transformation de variables
-  : Créer un tableau de contingence

- **Description des données** :
  -  : Statistiques descriptives
  -  : Test de normalité
  -  : Tableau après application du filtre
- **Visualisation des données** :
  -  : Histogrammes
  -  : Nuages de points
- **Analyse des données** :
  -  : Analyse en Composantes principales (ACP)
  -  : Classification k-means
  -  : Classification k-means floue
  -  : Classification Ascendante Hiérarchique (CAH)
- **Modélisation des données** :
  -  : Régression linéaire
  -  : ANOVA
  -  : Régression logistique
  -  : Régression PLS
- **Analyse de séries temporelles** :
  -  : ARIMA
  -  : Transformation de séries temporelles
  -  : Tests de tendance de Mann-Kendall
- **Analyse de données sensorielles** :
  -  : Variance par classe

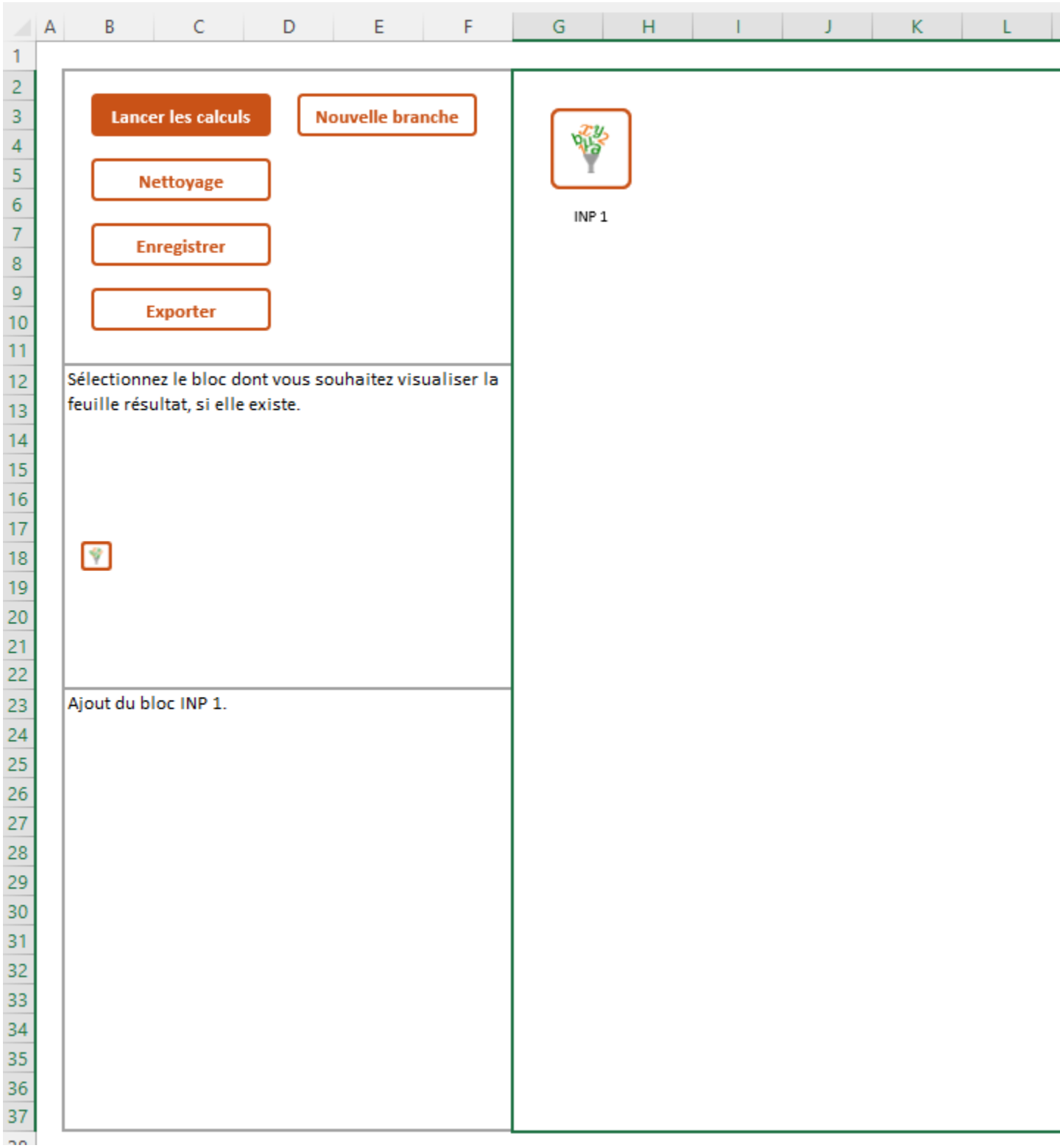
-  : Analyse de données CATA

- **Text mining** :
  -  : Nettoyer les données textuelles
  -  : Extraction de caractéristique

Cette liste est amenée à évoluer.

Écran d'une filière

L'écran d'une filière se présente comme sur la figure ci-dessous.



Voici le détail des différentes parties de l'écran :

- **Partie 1 (en haut à gauche)** : elle contient différents boutons dont voici le détail :

- **Lancer les calculs** : cliquez sur ce bouton si vous souhaitez lancer tous les calculs de la filière. Chaque bloc, depuis les données d'entrée jusqu'aux analyses finales, va être lancé afin de pouvoir visualiser le ou les résultats finaux. Si une analyse ne se déroule pas correctement alors les calculs s'arrêtent et l'utilisateur est averti. Il est possible de modifier les données d'entrées extérieures à la filière (les données d'entrées des blocs relatifs aux données d'entrées de la filière). Cette modification sera prise en compte pour les nouveaux calculs. Cependant, il peut arriver que ce changement impacte le nombre de colonnes de certains tableaux de sortie d'analyses. Il vous sera alors demandé de reparamétrer les données d'entrées des analyses concernées.

- **Nettoyage** : cliquez sur ce bouton si vous souhaitez nettoyer la partie 4 afin de repartir d'une feuille blanche. Attention, tous les blocs (données d'entrée et analyses) existants seront supprimés.

- **Enregistrer** : cliquez sur ce bouton si vous souhaitez enregistrer votre filière. Un fichier .wkf est créé dans le répertoire utilisateur XLSTAT. Il contient la structure et les paramètres de votre filière mais pas les données d'entrée. À la fermeture d'un classeur Excel contenant une ou plusieurs filières, il vous sera demandé si vous souhaitez sauver vos filières. En effet, une fois le classeur fermé, toute filière non sauvée sera perdue. Pour partager votre filière avec une autre personne, il faut exporter la filière concernée en cliquant sur le bouton **Exporter**.

- **Exporter** : cliquez sur ce bouton si vous souhaitez exporter votre filière. Une fenêtre va s'ouvrir afin de sélectionner le répertoire d'export et éventuellement modifier le nom du fichier exporté tout en gardant l'extension .wkf. Vous pourrez ainsi partager ce fichier avec une autre personne qui pourra visualiser et modifier la filière après l'avoir importée (voir Importer une filière). Il contient la structure et les paramètres de votre filière mais aussi les données d'entrée.

- **Nouvelle branche** : cliquez sur ce bouton si vous souhaitez créer une nouvelle branche. Un nouveau bloc pour les données d'entrée va apparaître dans la partie 4. Il sera sur la gauche et au premier emplacement disponible. Sa fenêtre de paramétrage va également s'afficher et vous pourrez lui succéder de nouvelles analyses ou des analyses déjà existantes en le connectant à une branche.

- **Partie 2 (au milieu à gauche)** : elle contient une reproduction mignature de la filière de la partie 4 et permet de visualiser rapidement les données d'entrée (si elles existent) ou la feuille résultat d'une analyse (si elle existe). Il suffit de cliquer sur la mignature du bloc cible.

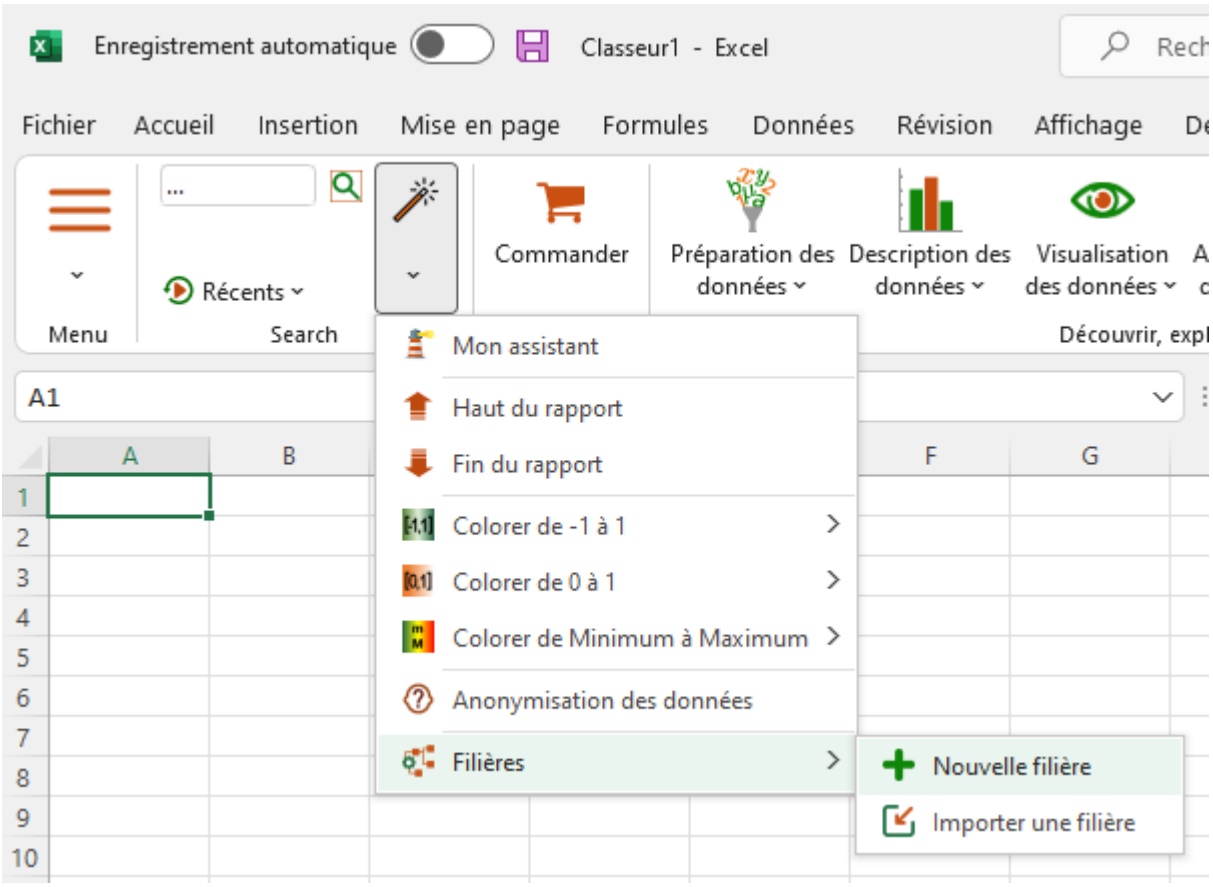
- **Partie 3 (en bas à gauche)** : elle contient des commentaires qui se mettent à jour automatiquement et renseignent sur les actions effectuées.

- **Partie 4 (à droite)** : elle contient la filière avec tous ses blocs, depuis lesquels il est possible d'effectuer certaines actions.

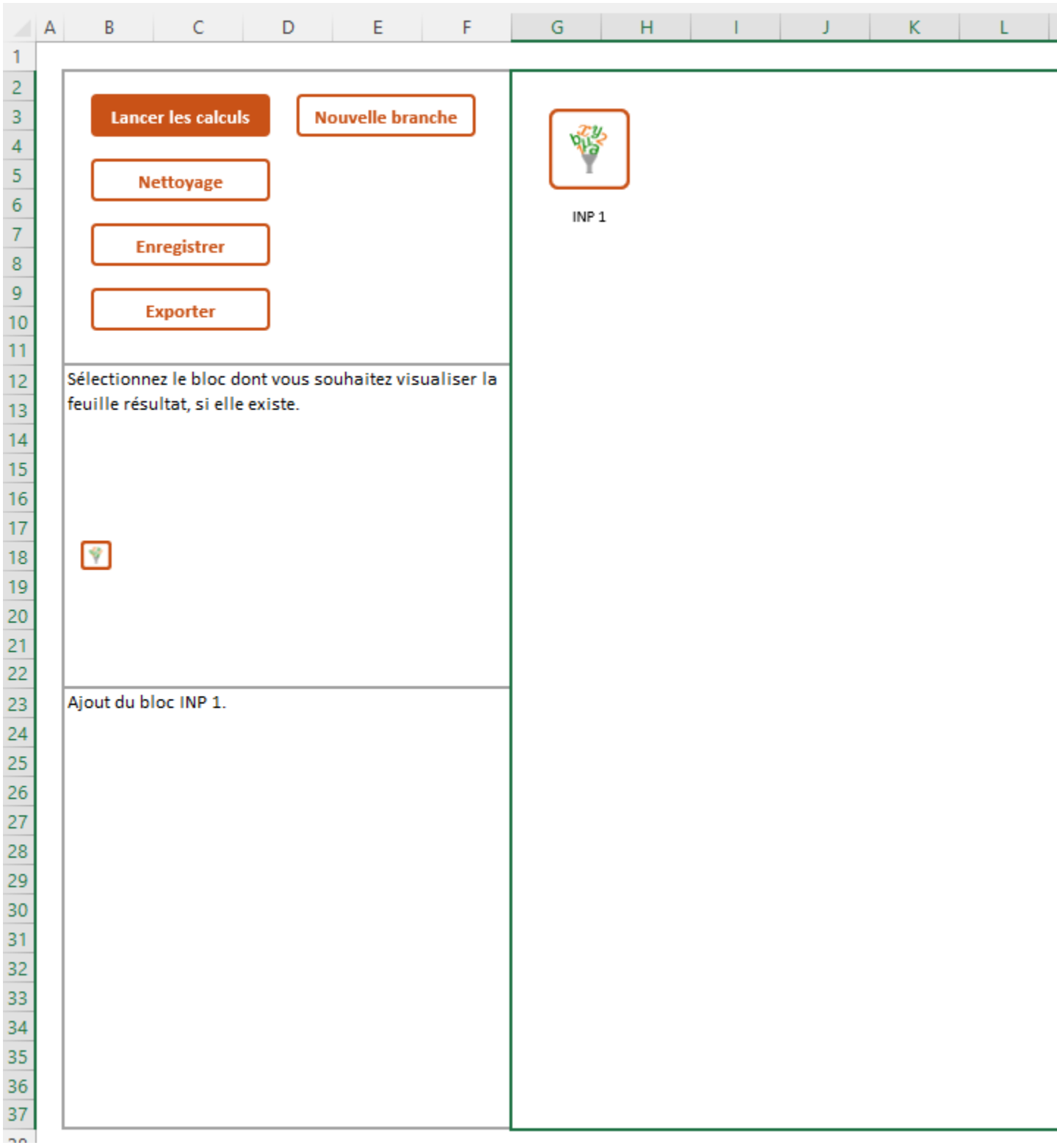
Création d'une nouvelle filière

Affichage de l'espace de travail d'une nouvelle filière

Pour construire un nouvelle filière, il faut commencer par afficher l'écran associé, présenté précédemment dans la partie Écran d'une filière. Pour cela, lancez XLSTAT et allez chercher l'outil Filières dans le ruban comme montré dans la figure ci-dessous.



Une nouvelle feuille Excel va ainsi être créée avec l'espace de travail de la nouvelle filière à construire.

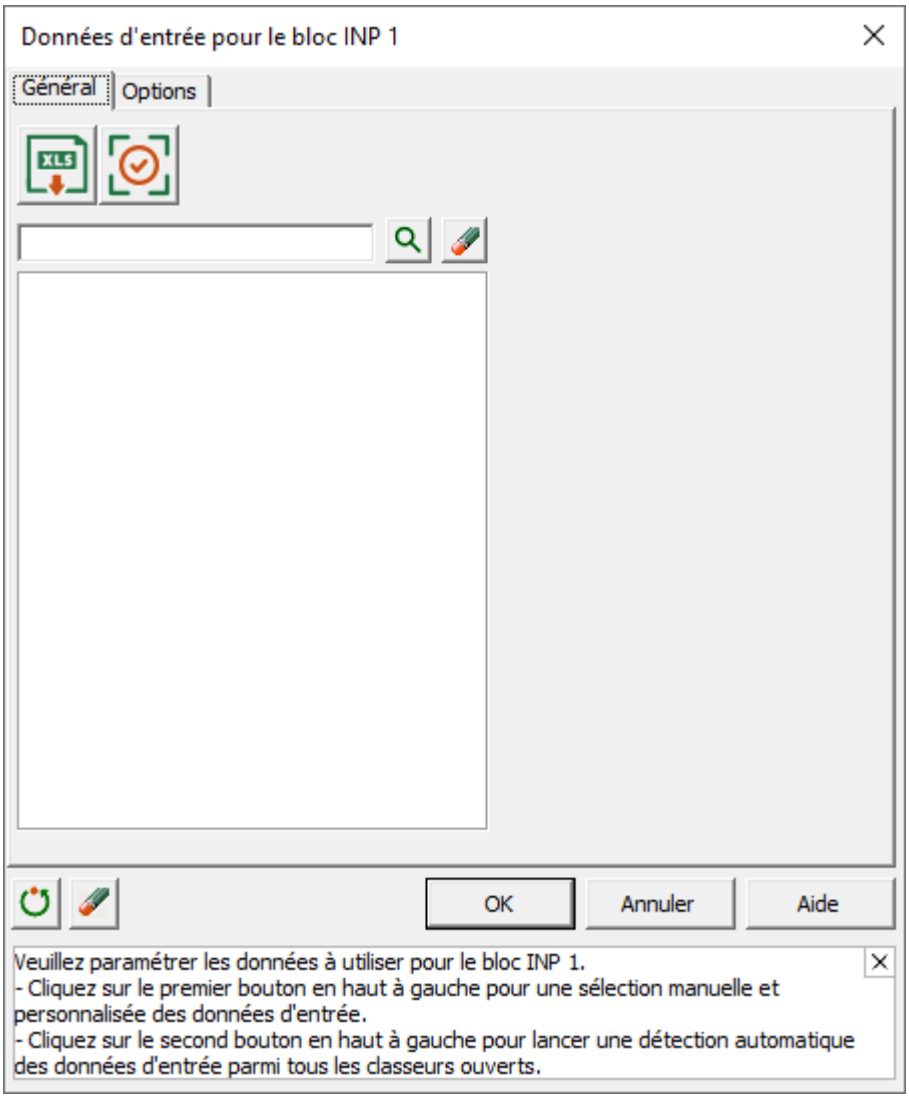


Par défaut, un premier bloc pour les données d'entrée s'affiche ainsi qu'une boîte de dialogue pour son paramétrage.

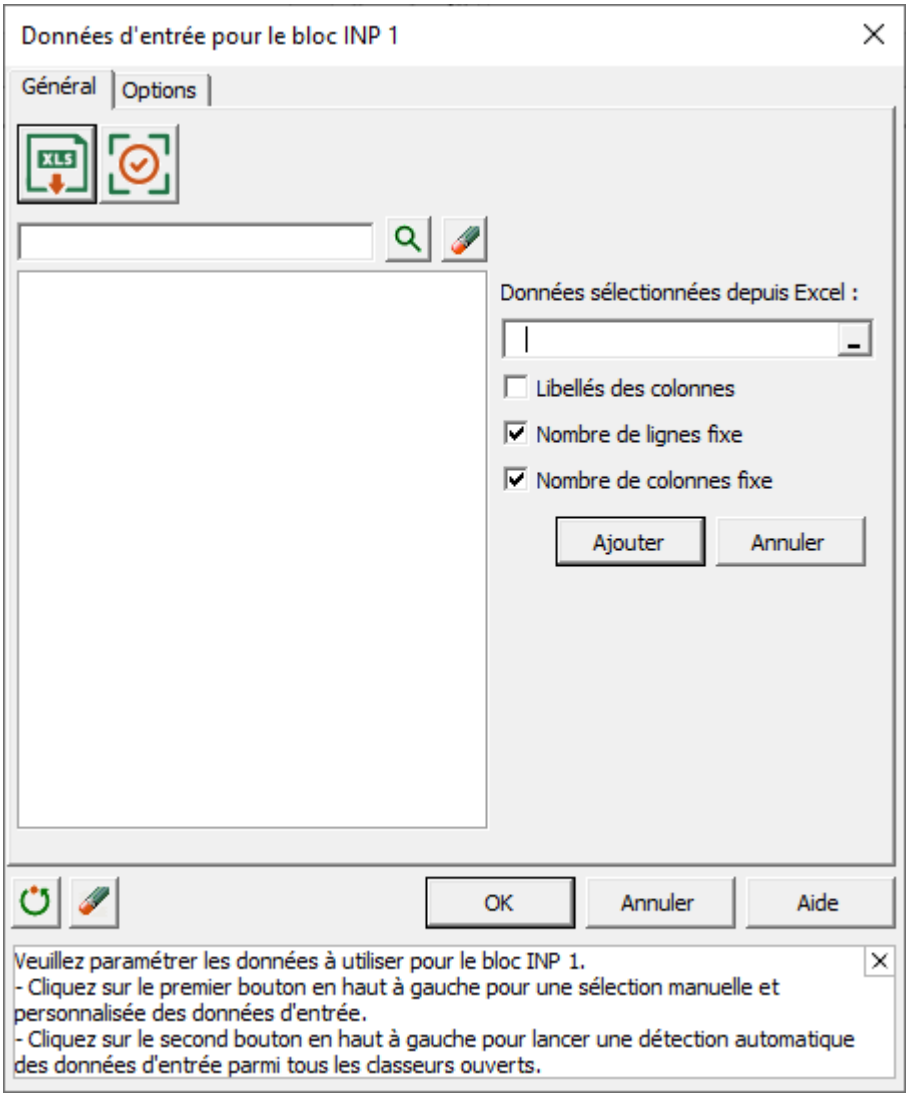
Paramétrage d'un premier bloc pour les données d'entrée

Nous allons détailler le contenu de la boîte de dialogue de paramétrage pour les données d'entrée.

Onglet **Général** :



: cliquez sur ce bouton si vous souhaitez sélectionner les données d'entrée manuellement parmi tous les classeurs ouverts.

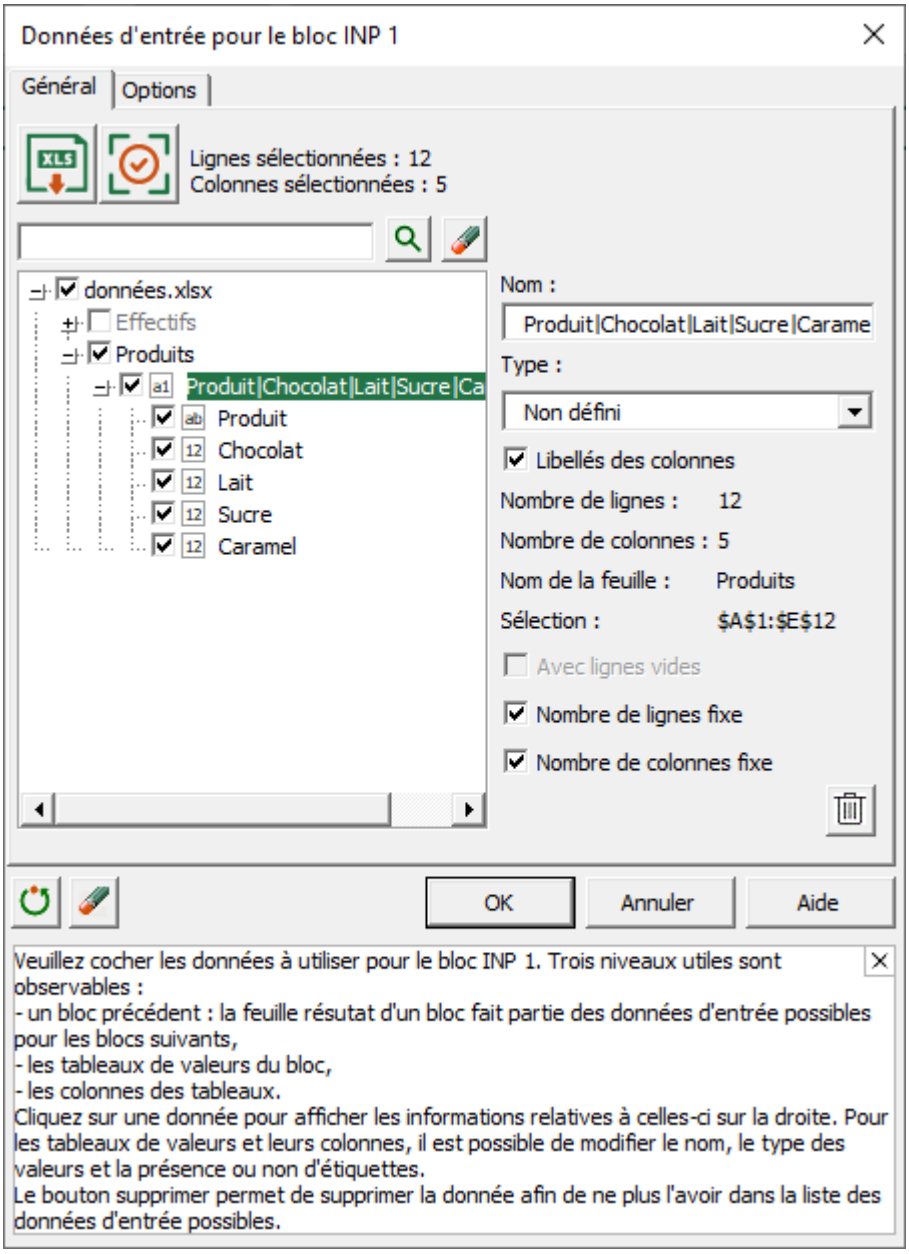


Commencez par sélectionner un premier tableau de données dans une feuille d'un classeur Excel ouvert. Vous pouvez ensuite mettre à jour les informations qui suivent :

- **Libellés des colonnes** : si coché alors cela signifie que les colonnes ont des libellés.
- **Nombre de lignes fixe** : si coché alors cela signifie que les nouvelles lignes ajoutées ultérieurement à la fin du tableau Excel sélectionné seront automatiquement ajoutées aux données du bloc. Toute insertion ou suppression de lignes à l'intérieur du tableau Excel sera automatiquement prise en compte, que la case soit cochée ou non.
- **Nombre de colonnes fixe** : si coché alors cela signifie que les nouvelles colonnes ajoutées ultérieurement à la fin du tableau Excel sélectionné seront automatiquement ajoutées aux données du bloc. Toute insertion ou suppression de colonnes à l'intérieur du tableau Excel sera automatiquement prise en compte, que la case soit cochée ou non. Attention, une suppression de colonnes utilisées par la suite dans la filière va entrainer une réinitialisation des blocs impactées avec suppression de leur feuille résultat si elle existe.

Cliquez sur **Ajouter** pour valider votre choix ou **Annuler**.


: cliquez sur ce bouton si vous souhaitez détecter automatiquement les données d'entrée parmi tous les classeurs ouverts. Cela va ajouter les données détectées dans la boîte de dialogue. Vous pouvez contrôler, modifier ou supprimer les données automatiquement chargées.



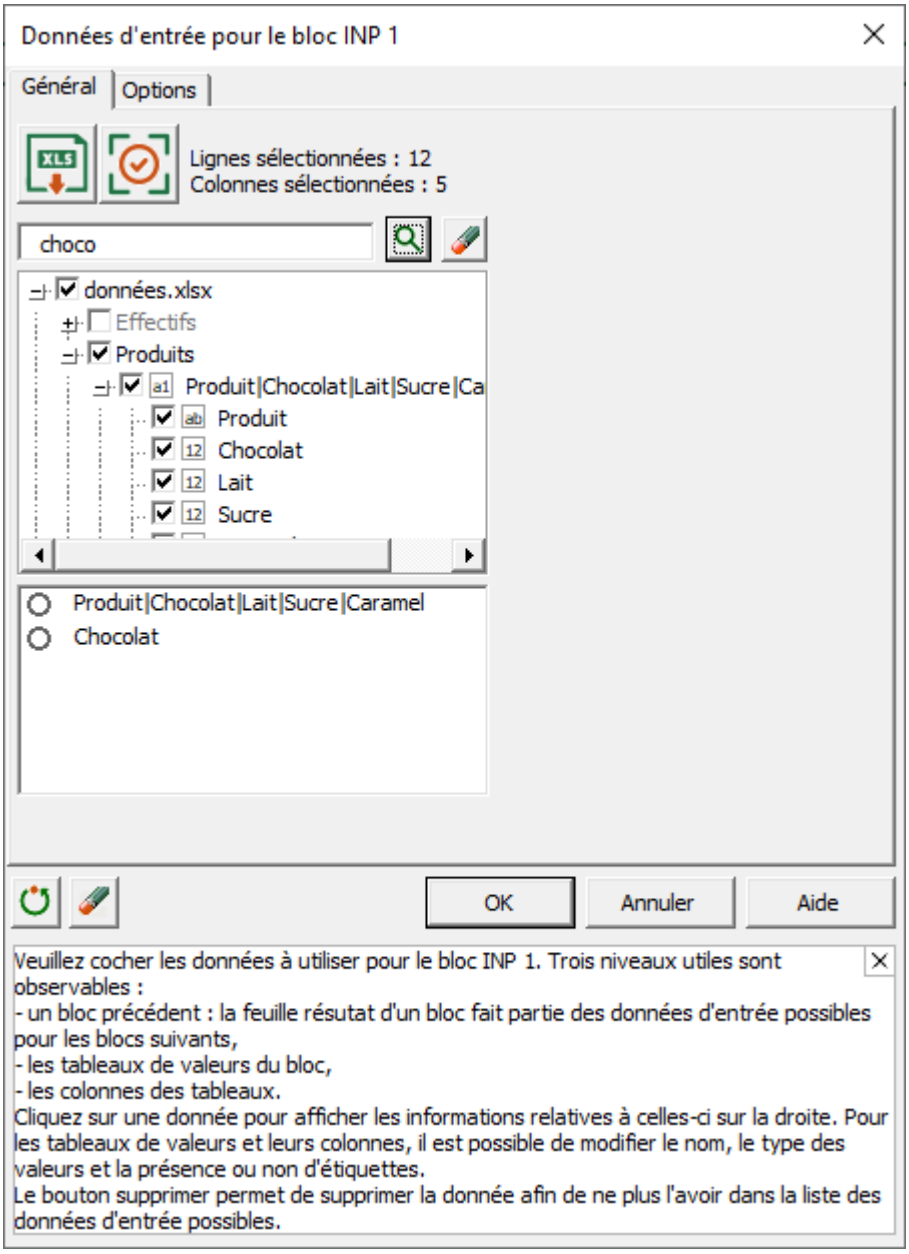
Nous pouvons observer quatre niveaux utiles :

- **Classeur** avec le nom du classeur où se trouve la plage de données.
- **Feuille** avec le nom de la feuille où se trouve la plage de données.
- **Tableau** avec les libellés des colonnes de la plage de données (ou la première et la dernière cellule si aucun libellé de colonne n'est détecté).
- **Colonne** avec le le libellé de la colonne concernée (ou la première et la dernière cellule si aucun libellé de colonne n'est détecté).

Il est possible de cliquer sur un des éléments de l'arborescence afin d'obtenir et éventuellement modifier certaines informations :

- **Nom** : nom de l'élément dans l'arborescence. Celui-ci peut être modifié.
- **Type** : type de l'élément : non défini, quantitatif ou qualitatif. Celui-ci peut être modifié.
- **Libellés des colonnes** : si coché alors cela signifie que l'élément à un libellé si c'est une colonne ou plusieurs si c'est un tableau.
- **Nombre de lignes**
- **Nombre de colonnes**
- **Nom de la feuille**
- **Sélection** : première et dernière cellules de la plage de données.
- **Avec lignes vides** : si coché alors cela signifie qu'il y a des lignes vides dans les données.
- **Nombre de lignes fixe** : si coché alors cela signifie que les nouvelles lignes ajoutées ultérieurement à la fin de la plage de données relative à l'élément sélectionné seront automatiquement ajoutées à celle-ci. Toute insertion ou suppression de lignes à l'intérieur de cette même plage de données sera automatiquement prise en compte, que la case soit cochée ou non. Cette information est modifiable seulement au niveau tableau.
- **Nombre de colonnes fixe** : si coché alors cela signifie que les nouvelles colonnes ajoutées ultérieurement à la fin de la plage de données relative à l'élément sélectionné seront automatiquement ajoutées aux données du bloc. Toute insertion ou suppression de colonnes à l'intérieur de cette même plage de données sera automatiquement prise en compte, que la case soit cochée ou non. Cette information est modifiable seulement au niveau tableau. Attention, une suppression de colonnes utilisées par la suite dans la filière va entrainer une réinitialisation des blocs impactées avec suppression de leur feuille résultat si elle existe.
-  : bouton afin de supprimer l'élément sélectionné dans l'arborescence.

Une barre de recherche se situe au-dessus de l'arboresence. Elle permet de retrouver facilement un élément à l'intérieur de celle-ci grâce à l'affichage qui apparaît en-dessous de l'arborescence comme dans la figure ci-dessous.



Et pour finir :

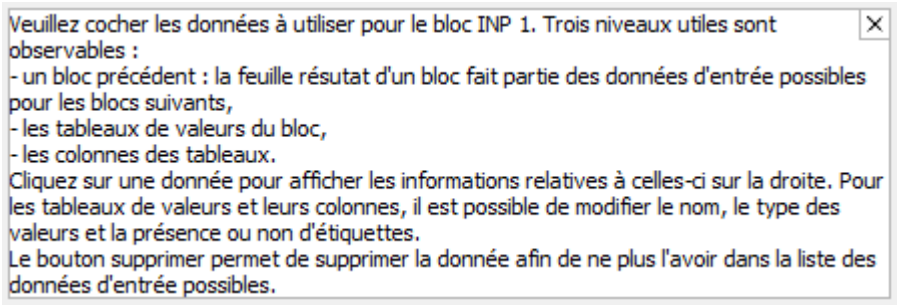
OK : cliquez sur ce bouton pour valider et sauvegarder vos paramètres.

Annuler : cliquez sur ce bouton pour fermer la boîte de dialogue sans sauvegarder les paramètres.

Aide : cliquez sur ce bouton pour afficher l'aide relative aux filières dans XLSTAT.

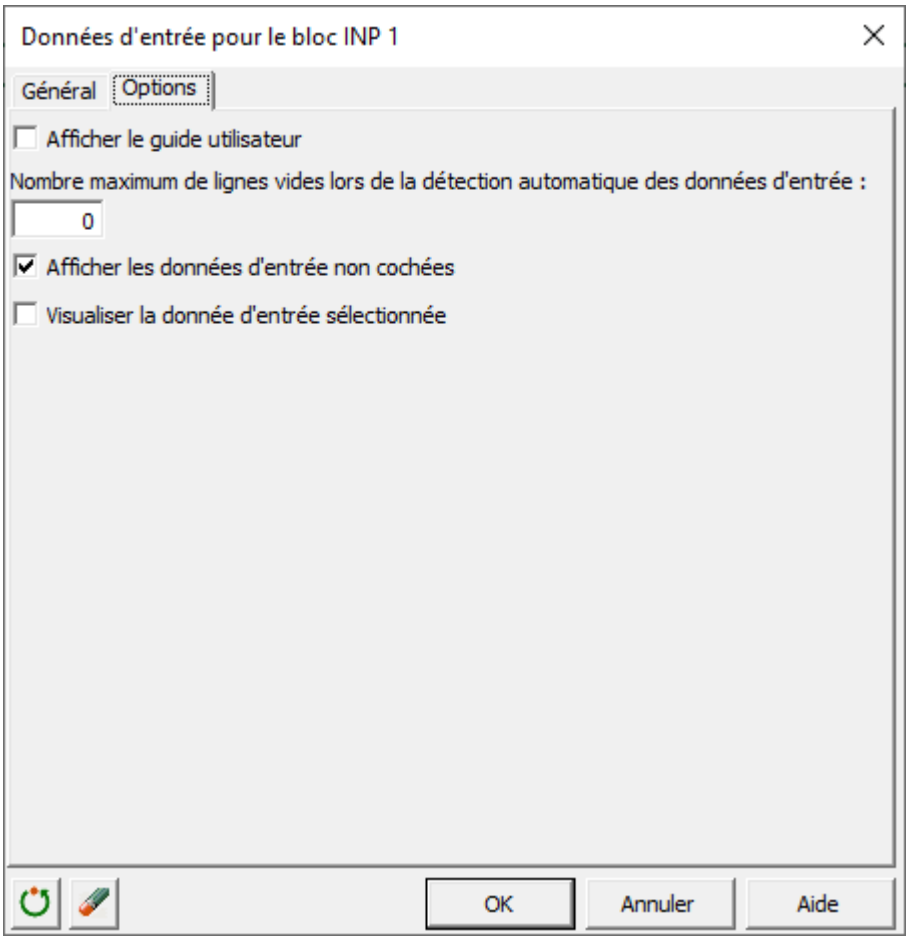
: cliquez sur ce bouton pour réinitialiser la boîte de dialogue avec les paramètres par défaut.

: cliquez sur ce bouton pour effacer les sélections de données de la boîte de dialogue.



: cet encart sert de guide utilisateur quant aux actions à mener. Il peut être caché en cliquant sur la croix en haut à droite de celui-ci ou via les options.

Onglet **Options** :



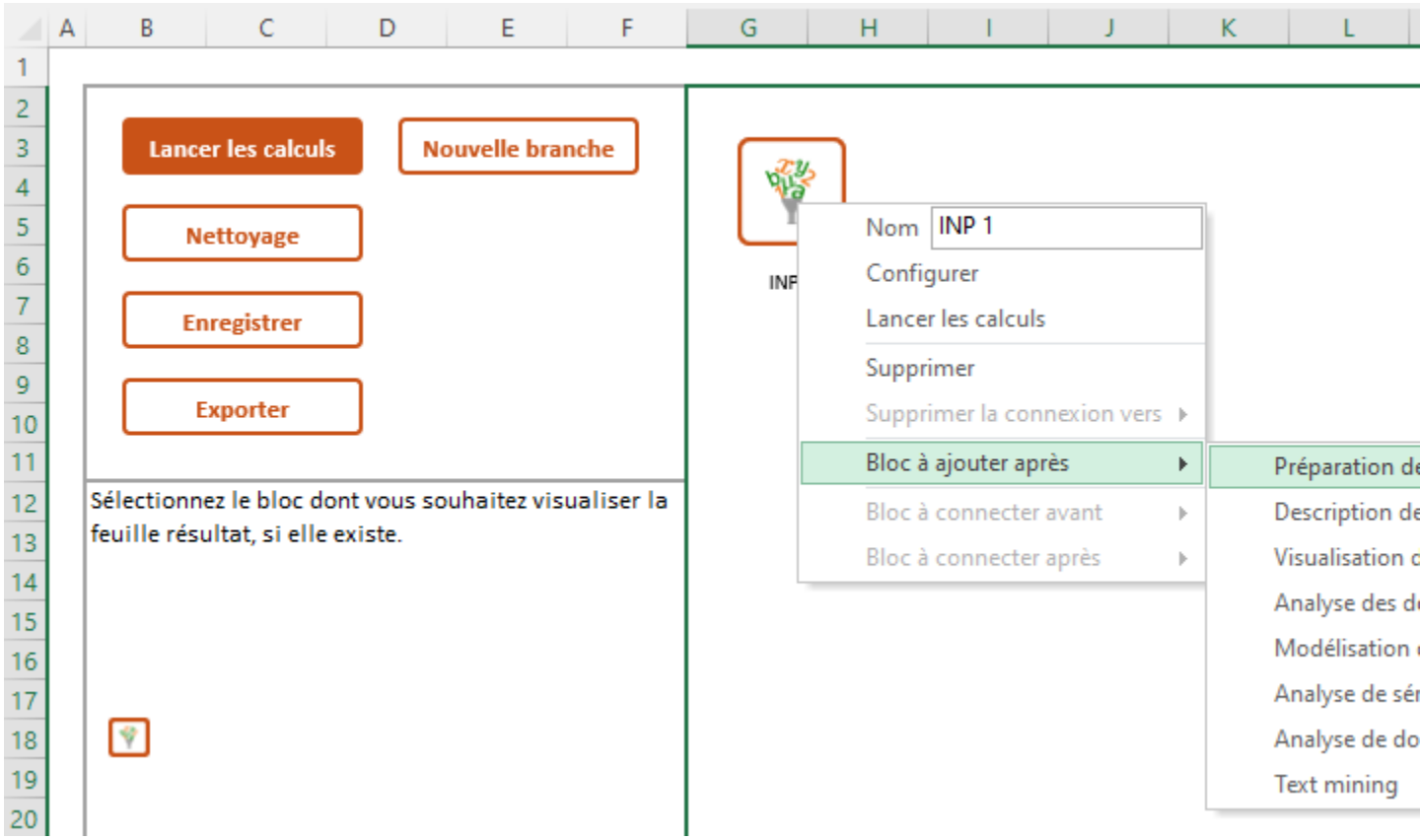
- **Afficher le guide utilisateur** : cochez cette case si vous souhaitez afficher le guide utilisateur qui se trouve en bas de la boîte de dialogue.
- **Nombre maximum de lignes vides lors de la détection automatique des données d'entrée** : si vous sélectionnez vos données d'entrée à partir de l'outil de détection automatique, vous pouvez préciser le nombre maximun de lignes vides toléré. Si celui-ci est à 0 (valeur par défaut) alors aucune ligne vide ne sera acceptée.
- **Afficher les données d'entrée non cochées** : cochez cette case si vous souhaitez afficher les données d'entrée non cochées. Dans l'onglet Général, surtout si vous choisissez de faire une détection automatique, il est fort probable que toutes les données ne vous intéressent pas. Une fois votre sélection faite, vous pouvez choisir de ne plus

afficher les autres, tout en les gardant en mémoire. Nous avons vu plus haut qu'il est aussi possible de les supprimer.

- **Visualiser la donnée d'entrée sélectionnée** : cochez cette case si vous souhaitez visualiser la donnée d'entrée sélectionnée dans l'arborescence de l'onglet Général.

Ajout d'un nouveau bloc

Une fois le bloc de données d'entrée ajouté, vous pouvez cliquer sur celui-ci afin d'afficher le menu des différentes actions possibles à partir de ce bloc. Celles-ci seront détaillées dans la partie Actions sur une filière. Cependant, nous allons déjà présenter celle concernant l'ajout de bloc.

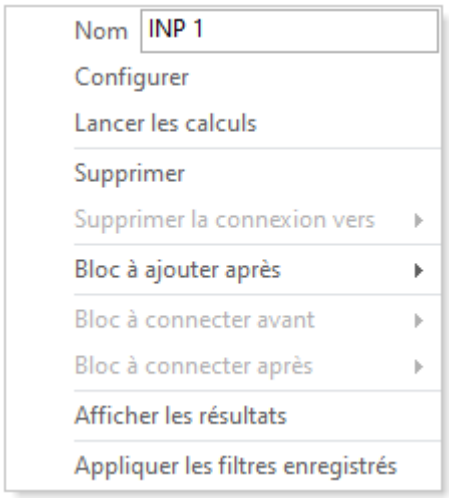


Dans le menu, allez sur "Bloc à ajouter après", puis cliquez sur l'analyse qui vous intéresse. Un nouveau bloc va être ajouté à la suite de la filière et la boîte de dialogue associée va apparaître. Il vous suffit de paramétrer celle-ci comme voulu. Pour accéder à la documentation de celle-ci, vous pouvez cliquer sur le nom de l'analyse cible dans la partie Différentes analyses utilisables.

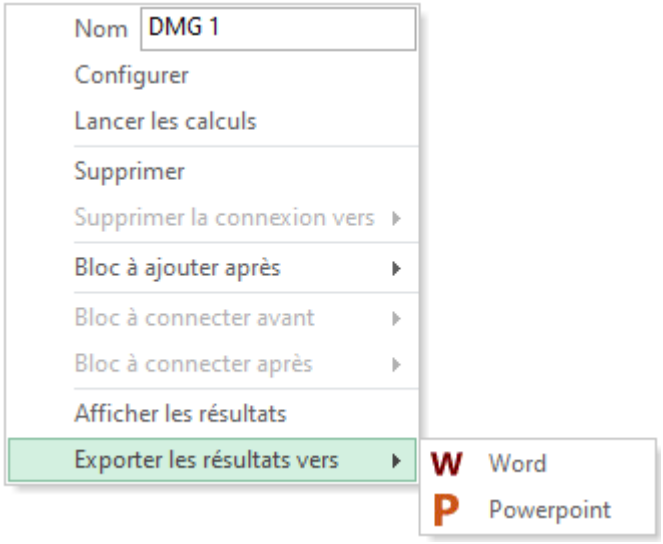
Vous pouvez ainsi ajouter autant de blocs que vous le souhaitez. Pour rappel, les Statistiques descriptives, les Histogrammes et les Nuages de points ne peuvent pas avoir de bloc qui les succède.

Actions sur un bloc d'une filière

Menu pour un bloc de données d'entrée :



Menu pour un bloc d'analyse :

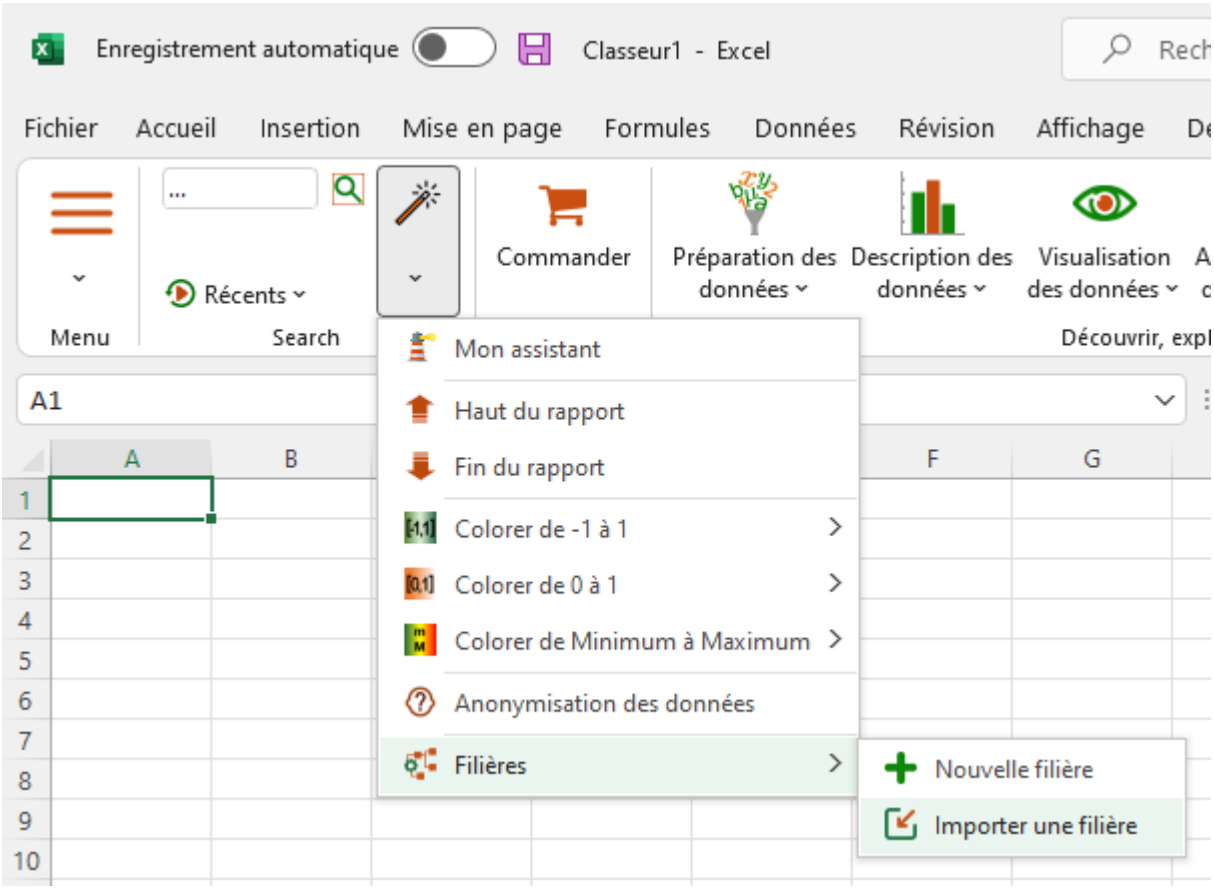


- **Nom** : vous pouvez modifier le nom d'un bloc via cette zone de texte. Celui-ci se mettra à jour après avoir cliqué sur Entrée.
- **Configurer** : cliquez sur cet élément si vous souhaitez afficher la boîte de dialogue pour le paramétrage du bloc. Si vous configurez un bloc alors les blocs connectés ensuite peuvent être impactés. Dans ce cas, les feuilles résultats des blocs impactés seront supprimées et il vous sera demandé de relancer les calculs si vous souhaitez les mettre à jour. Dans une filière, un connecteur de couleur vert signifie que le bloc vers lequel celui-ci pointe possède une feuille résultat qui peut être visualisée.
- **Lancer les calculs** : cliquez sur cet élément si vous souhaitez lancer les calculs jusqu'au bloc. Tous les blocs d'analyses appartenant à une même branche, du plus à gauche et jusqu'à celui-ci auront leurs résultats mis à jour.

- **Supprimer** : cliquez sur cet élément si vous souhaitez supprimer le bloc. Tous les blocs suivants, avec une connexion directe ou non, seront également supprimés.
- **Supprimer la connexion vers** : sélectionnez cet élément et cliquez sur le nom du bloc directement connecté après dont vous souhaitez supprimer la connexion. Si ce bloc suivant se retrouve sans bloc précédent alors il est supprimé et tous les blocs suivants, avec une connexion directe ou non, seront également supprimés.
- **Bloc à ajouter après** : sélectionnez cet élément et cliquez sur le bloc d'analyse que vous souhaitez ajouter après.
- **Bloc à connecter avant** : sélectionnez cet élément et cliquez sur le nom du bloc que vous souhaitez connecter avant. Si vous ne trouvez pas le bloc voulu alors il ne fait pas partie des blocs possibles.
- **Bloc à connecter après** : sélectionnez cet élément et cliquez sur le nom du bloc que vous souhaitez connecter après. Si vous ne trouvez pas le bloc voulu alors il ne fait pas partie des blocs possibles.
- **Afficher les résultat** : cet élément n'est actif dans le menu que pour les blocs qui ont une feuille résultat. Cliquez dessus si vous souhaitez afficher le résultat de l'analyse associée au bloc. Vous serez directement envoyé sur la feuille résultat. Les calculs ne seront pas relancés.
- **Appliquer les filtres enregistrés** : cet élément n'est actif dans le menu que pour les blocs de données d'entrée qui ont des données filtrées. Cliquez dessus si vous souhaitez appliquer, dans les feuilles concernées, les filtres utilisés pour ce bloc. Il est possible d'avoir plusieurs blocs de données d'entrée avec des données provenant de mêmes feuilles mais avec des filtres différents.
- **Exporter les résultats vers** : cet élément n'est actif dans le menu que pour les blocs qui ont une feuille résultat. Sélectionnez le pour exporter les résultats du bloc. Cliquez sur Word ou Powerpoint et une boîte de dialogue s'affichera afin de choisir les éléments de la feuille résultat à exporter.

Importer une filière

Nous avons vu qu'il est possible d'exporter une filière via le bouton **Exporter** à gauche dans l'espace de travail de celle-ci. Cela vous permettra donc d'importer ultérieurement une filière exportée. Le bouton d'import se trouve dans le ruban au niveau de l'outil Filières comme montré dans la figure ci-dessous.



Plusieurs feuilles Excel vont être ajoutées au classeur, celles relatives au jeu de données utilisé pour la filière importée et une autre avec l'espace de travail de celle-ci. Il vous sera demandé si vous souhaitez lancer les calculs ou non.

Exemples

Des exemples d'utilisation de l'outil Filières sont disponibles sur le Centre d'aide XLSTAT :

http://www.xlstat.com/demo-wkff.htm

Préparation des données

Echantillonnage de données

Utilisez cet outil pour générer un sous-échantillon d'observations à partir d'un jeu de données univariées ou multivariées.

Dans cette section :

Description

L'échantillonnage est l'une des techniques fondamentales. La génération d'échantillons permet notamment :

- de tester une hypothèse sur un échantillon, puis de la valider sur un autre ;
- d'obtenir des tableaux d'une taille plus petite tout en gardant des propriétés du tableau d'origine.

Afin de répondre à différentes situations, plusieurs méthodes ont été proposées. XLSTAT propose les méthodes suivantes pour générer un échantillon de N observations à partir d'un tableau de M lignes :

N premières lignes : l'échantillon obtenu est constitué des N premières lignes du tableau initial. Cette méthode n'est à utiliser que si l'on est sûr que les données n'ont pas été triées suivant un critère qui pourrait introduire un biais pour l'analyse.

N dernières lignes : l'échantillon obtenu est constitué des N dernières lignes du tableau initial. Cette méthode n'est à utiliser que si l'on est sûr que les données n'ont pas été triées suivant un critère qui pourrait introduire un biais pour l'analyse.

N tous les s, début à k : l'échantillon est obtenu en prenant N lignes toutes les s lignes, en commençant à la ligne k.

Aléatoire sans remise : des observations sont choisies au hasard et ne peuvent figurer qu'une seule fois dans l'échantillon.

Bootstrap (aléatoire avec remise) : des observations sont choisies au hasard et peuvent figurer plusieurs fois dans l'échantillon.

Systématique à départ aléatoire : à partir de la j-ième observation du tableau initial, une observation est extraite pour l'échantillon toutes les k observations. j est choisi au hasard parmi un nombre de possibilités dépendant de la taille du tableau initial et de la taille de l'échantillon final. k est déterminé de telle sorte que les observations extraites soient le plus possible espacées.

Systématique centré : les observations sont choisies de façon régulière aux centres de N séquences d'observations de même longueur k.

Aléatoire stratifié (1) à un élément par strate : des lignes sont choisies de façon aléatoire à l'intérieur de N séquences d'observations de même longueur, où N est déterminé en divisant le nombre d'observations par la taille d'échantillon souhaitée.

Aléatoire stratifié (2) : des lignes sont choisies de façon aléatoire à l'intérieur de N strates définies par l'utilisateur. Dans chaque strate, le nombre d'observations échantillonnées est proportionnel à la fréquence de la strate.

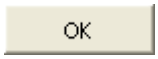
Aléatoire stratifié (3) : des lignes sont choisies de façon aléatoire à l'intérieur de N strates définies par l'utilisateur. Dans chaque strate, le nombre d'observations échantillonnées est proportionnel à une fréquence définie par l'utilisateur.

Défini par l'utilisateur : une variable indique la fréquence des observations dans l'échantillon à générer.

Echantillons d'apprentissage et de test : les données sont divisées en deux – un échantillon d'apprentissage et un échantillon de test. Les lignes de chaque échantillon sont tirées aléatoirement du jeu de données initial. La taille de l'échantillon d'apprentissage est définie en nombre de lignes.

Echantillons d'apprentissage et de test (%) : les données sont divisées en deux – un échantillon d'apprentissage et un échantillon de test. Les lignes de chaque échantillon sont tirées aléatoirement du jeu de données initial. La taille de l'échantillon d'apprentissage est définie en pourcentage du nombre de lignes du tableau initial.

Boîte de dialogue



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont vous chargez les données dans XLSTAT. Si le bouton possède une icône de souris, XLSTAT vous permet de faire une sélection à la souris des données. Si le bouton possède une icône de liste, XLSTAT vous permet de

sélectionner les données à partir d'une liste. Si le bouton possède une icône orange, des boutons complémentaires avec un point d'interrogation  sont affichés pour vous permettre d'importer les données à partir de fichiers.

Données : sélectionnez le tableau des données sur sur la feuille Excel.

Echantillonnage : choisissez la méthode d'échantillonnage (voir la section description pour plus d'information).

Taille d'échantillon : entrez la taille de l'échantillon à générer.

Strates : cette option est disponible pour les échantillonnages stratifiés (2) et (3). Sélectionnez dans ce champ une colonne indiquant à quelle strate correspond chaque observation.

Poids de chaque strate : cette option est disponible pour l'échantillonnage stratifié (3). Sélectionnez un tableau à deux colonnes, la première contenant l'identifiant de la strate, et la seconde le poids de la strate dans l'échantillon final. Quelque soit l'unité des poids (effectif, fréquence, pourcentage), XLSTAT standardise les poids de manière à ce que la somme soit égale à la taille d'échantillon demandée.

Effectifs : si l'échantillonnage est « défini par l'utilisateur », vous devez alors sélectionner une colonne de données dans ce champ. La colonne doit contenir le nombre de fois où chaque observation doit être sélectionnée.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des variables : activez cette option si la première ligne des données sélectionnées (données et libellés des observations) contient un libellé.

Libellés des observations : activez cette option si vous voulez utiliser des libellés d'observations disponibles sur une feuille Excel pour l'affichage des résultats. Si l'option « Libellés des variables » est activée, la première cellule de la sélection doit comprendre un en-tête. Si vous n'activez pas cette option, des libellés seront automatiquement créés (Obs1, Obs2, …).

Afficher l'en-tête du rapport : désactivez cette option si vous souhaitez que le tableau échantillonné commence dès la première ligne de la feuille Excel (cas d'une sortie dans une feuille ou un classeur), et non après l'en- tête du rapport. Vous pourrez ainsi sélectionner les variables de ce tableau par colonnes.

Mélanger : activez cette option si vous souhaitez que les données de sortie soient permutées de manière aléatoire. Si cette option n'est pas activée, les données échantillonnées respectent l'ordre des données de départ.

Afficher côte à côte : activez cette option pour afficher les échantillons générés les uns à côté des autres.

Bibliographie

Cochran W.G. (1977). Sampling Techniques. Third edition. John Wiley and Sons, New York.

Hedayat A.S. and Sinha B.K. (1991). Design and Inference in Finite Population Sampling. John Wiley and Sons, New York.

Echantillonnage dans une distribution

Utilisez cet outil pour générer un échantillon de données à partir d'une distribution théorique continue ou discrète, ou à partir d'un échantillon existant.

Dans cette section :

Description

Dans le cas où l'échantillon est généré à partir d'une distribution théorique, vous devez choisir la loi de probabilité, puis pour certaines d'entre elles, vous devez ensuite entrer la valeur des paramètres.

XLSTAT permet l'utilisation des lois suivantes :

- Arcsinus (a) : la densité de cette loi (dérivée de la loi Bêta de type I) est donnée par :

$$f(x) = \frac{\sin(\pi\alpha)}{\pi x}(\frac{x}{1-x})^{\alpha-1}, \quad \text{avec } 0 < \alpha < 1, x \in [0,1]$$

On a $E(X) = \alpha$ et $V(X) = \alpha(1-\alpha)/2$

- Bernoulli (p) : la densité de cette loi est donnée par :

$$P(X=1) = p, P(X=0) = 1-p, \quad \text{avec } p \in [0,1]$$

On a $E(X) = p$ et $V(X) = p(1-p)$

La loi de Bernoulli, du nom du mathématicien suisse Jacob Bernoulli (1654-1705), permet de décrire les phénomènes aléatoires binaires où seuls deux événements peuvent survenir avec des probabilités respectives de p et $1-p$.

- Bêta (a, b) : la densité de cette loi (aussi appelée Bêta de type I) est donnée par :

$$f(x) = \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}$$

$$\text{avec} \quad \alpha, \beta > 0, x \in [0,1] \text{ et } B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

On a $E(X) = \alpha/(\alpha+\beta)$ et $V(X) = \alpha\beta/[(\alpha+\beta+1)(\alpha+\beta)^2]$

- Bêta4 (α, β, c, d) : la densité de cette loi est donnée par :

$$f(x) = \frac{1}{B(\alpha, \beta)} \frac{(x-c)^{\alpha-1}(d-x)^{\beta-1}}{(d-c)^{\alpha+\beta-1}}, \quad \text{avec } \alpha, \beta > 0, x \in [c, d]$$

$$c, d \in \mathbb{R} \text{ et } B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

On a $E(X) = \frac{c+(c-d)\alpha}{(\alpha+\beta)}$ et $V(X) = \frac{(c-d)^2\alpha\beta}{(\alpha+\beta+1)(\alpha+\beta)^2}$

Pour la loi Bêta de type I, la distribution est dans l'intervalle $[0, 1]$. La loi Bêta4 est obtenue par un simple changement de variable de la loi Bêta de type I de telle sorte que la distribution soit sur l'intervalle [c, d].

- Binomiale (n, p) : la densité de cette loi est donnée par :

$$P(X = x) = C_n^x p^x (1-p)^{n-x}, \quad \text{avec } n, x \in \mathbb{N}, n > 0, x \in [0, n], p \in [0, 1]$$

On a $E(X) = np$ et $V(X) = np(1-p)$

n est le nombre d'essais, et p la probabilité de succès. La loi binomiale est la loi du nombre de succès pour n essais, sachant que la probabilité de succès vaut p. La loi binomiale peut être vue comme la loi de n tirages dans une loi de Bernoulli.

- Binomiale négative (n, p) de type I : la densité de cette loi est donnée par :

$$P(X = x) = C_{n+x-1}^{x-1} p^n (1-p)^x, \quad \text{avec } n, x \in \mathbb{N}, n > 0, p \in [0, 1]$$

On a $E(X) = n(1-p)/p$ et $V(X) = n(1-p)/p^2$

n est le nombre de succès et p la probabilité de succès. La loi binomiale négative de type I est la loi du nombre de tirages x sans succès nécessaires avant d'avoir obtenus n succès.

- Binomiale négative (k, p) de type II : la densité de cette loi est donnée par :

$$P(X = x) = \frac{\Gamma(k+x)p^x}{x!\Gamma(k)(1+p)^{k+x}}, \quad \text{avec } x \in N, k, p > 0$$

On a $E(X) = kp$ et $V(X) = kp(p+1)$

La loi binomiale négative de type II permet de représenter des phénomènes discrets fortement hétérogènes. Lorsque k tend vers l'infini, la loi binomiale négative de type II tend vers une loi de Poisson de paramètre $\lambda = kp$.

- $Khi^2(df)$: la densité de cette loi est donnée par :

$$f(x) = \frac{(1/2)^{df/2}}{\Gamma(df/2)} x^{\frac{df}{2}-1} e^{-x/2}, \quad \text{avec } x > 0, df \in \mathbb{N}^*$$

On a $E(X) = df$ et $V(X) = 2df$

La loi du Khi² correspond à la loi de la somme des carrés de df lois normales centrées réduites (lois normales standard). Elle est très utilisée pour tester des hypothèses.

- Erlang (k, λ) : la densité de cette loi est donnée par :

$$f(x) = \lambda^k x^{k-1} \frac{e^{-\lambda x}}{(k-1)!}, \quad \text{avec } x \geq 0 \text{ et } k, \lambda > 0 \text{ et } k \in \mathbb{N}$$

On a $E(X) = k/\lambda$ et $V(X) = k/\lambda^2$

k est le paramètre de forme de la loi et λ est le paramètre de taux.

Cette distribution, développée par le scientifique danois A. K. Erlang (1878-1929) pour l'étude du trafic téléphonique, est utilisée de manière plus générale pour l'étude des files d'attente.

Remarque : lorsque $k = 1$, cette distribution est équivalente à la distribution exponentielle, et la loi Gamma à deux paramètres est une généralisation de la loi d'Erlang au cas où k est un réel et non un entier (par ailleurs on utilise le paramètre d'échelle $\beta = 1/\lambda$).

- Exponentielle (λ) : la densité de cette loi est donnée par :

$$f(x) = \lambda \exp(-\lambda x), \quad \text{avec } x > 0 \text{ et } \lambda > 0$$

On a $E(X) = 1/\lambda$ et $V(X) = 1/\lambda^2$

La loi exponentielle est souvent utilisée pour étudier la durée de vie en contrôle qualité.

- Fisher (df_1, df_2) : la densité de cette loi est donnée par :

$$f(x) = \frac{1}{xB(df_1/2, df_2/2)} \left(\frac{df_1 x}{df_1 x + df_2} \right)^{df_1/2} \left(1 - \frac{df_1 x}{df_1 x + df_2} \right)^{df_2/2}$$

$$\text{avec } x > 0 \text{ et } df_1, df_2 \in \mathbb{N}^*$$

On a $E(X) = df_2/(df_2 - 2)$ si $df_2 > 2$, et $V(X) = \frac{2df_2^2(df_1 + df_2 - 2)}{df_1(df_2-2)^2(df_2-4)}$ si $df_2 > 4$

La loi de Fisher, du nom du biologiste, généticien et statisticien Ronald Aylmer Fisher (1890-1962), correspond au rapport de deux lois du Khi². Elle est très utilisée pour tester des hypothèses.

- Fisher-Tippett (β, μ) : la densité de cette loi est donnée par :

$$f(x) = \frac{1}{\beta} \exp\left(-\frac{x-\mu}{\beta} - \exp\left(-\frac{x-\mu}{\beta} \right) \right), \quad \text{avec } \beta > 0$$

On a $E(X) = \mu + \beta\gamma$ et $V(X) = (\pi\beta)^2/6$ où γ est la constante de Euler-Mascheroni.

La loi de Fisher-Tippett, aussi appelée loi Log-Weibull, ou loi généralisée des valeurs extrêmes, est utilisée dans l'étude de phénomènes extrêmes. La loi de Gumbel est un cas particulier de la loi de Fisher-Tippett avec $\beta = 1$ et $\mu = 0$.

- Gamma (k, β, μ) : la densité de cette loi est donnée par :

$$f(x) = (x-\mu)^{k-1}\ \frac{e^{-(x-\mu)/\beta}}{\beta^k \Gamma(k)}, \quad \text{avec } x > \mu \text{ et } k, \beta > 0$$

On a $E(X) = \mu + k\beta$ et $V(X) = k\beta^2$

k est le paramètre de forme de la loi et β est le paramètre d'échelle.

- GEV (β, k, μ): la densité de cette loi est donnée par :

$$f(x) = \frac{1}{\beta}\left(1 + k\frac{x-\mu}{\beta}\right)^{-1/k-1} \exp\left(-\left(1 + k\frac{x-\mu}{\beta}\right)^{-1/k}\right), \quad \text{avec} \quad \beta > 0$$

On a $E(X) = \mu + \frac{\beta}{k}\Gamma(1+k)$ et $V(X) = \left(\frac{\beta}{k}\right)^2 \left(\Gamma(1+2k) - \Gamma^2(1+k)\right)$

La loi GEV (Generalized Extreme Values) est très utilisée en hydrologie pour modéliser les phénomènes de crues. k est classiquement compris entre -0.6 et 0.6.

- Gumbel : la densité de cette loi est donnée par :

$$f(x) = \exp(-x - \exp(-x))$$

On a $E(X) = \gamma$ et $V(X) = \pi^2/6$ où γ est la constante de Euler-Mascheroni (0.5772156649…).

La loi de Gumbel, du nom de Emil Julius Gumbel (1891-1966), est un cas particulier de la loi de Fisher-Tippett avec $\beta = 1$ et $\mu = 0$. Elle est utilisée dans l'étude de phénomènes extrêmes comme les précipitations ou les crues maximales et les magnitudes maximales de tremblement de terre.

- Logistique (μ, s) : la densité de cette loi est donnée par :

$$f(x) = \frac{e^{-\frac{(x-\mu)}{s}}}{s(1 + e^{-\frac{(x-\mu)}{s}})}, \quad \text{avec } s > 0$$

On a $E(X) = \mu$ et $V(X) = (\pi s)^2/3$

- Lognormale (μ, σ) : la densité de cette loi est donnée par :

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(\ln(x)-\mu)^2}{2\sigma^2}}, \quad \text{avec } x, \sigma > 0$$

On a $E(X) = exp(\mu + \sigma^2/2)$ et $V(X) = \left(exp(\sigma^2) - 1\right) exp(2\mu + \sigma^2)$

- Lognormale2 (m, s) : la densité de cette loi est donnée par :

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(\ln(x)-\mu)^2}{2\sigma^2}}, \text{ avec } x, \sigma > 0$$

On a :

$\mu = ln(m) - ln(1 + s^2/m^2)/2$ et $\sigma^2 = ln(1 + s^2/m^2)$

Et :

$E(X) = m$ et $V(X) = s^2$

Cette distribution est simplement une reparamétrisation de la loi Lognormale.

- Normale (μ, σ) : la densité de cette loi est donnée par :

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad \text{avec } \sigma > 0$$

On a $E(X) = \mu$ et $V(X) = \sigma^2$

- Normale standard : la densité de cette loi est donnée par :

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

On a $E(X) = 0$ et $V(X) = 1$

Cette loi est un cas particulier de la loi normale, avec $\mu = 0$ et $\sigma = 1$. Elle est aussi appelée loi normale centrée réduite.

- Pareto (a, b) : la densité de cette loi est donnée par :

$$f(x) = \frac{ab^a}{x^{a+1}}, \text{ avec } a, b > 0 \text{ et } x \geq b$$

On a $E(X) = ab/(a-1)$ et $V(X) = \frac{ab^2}{((a-1)^2(a-2))}$

La loi de Pareto, du nom de l'économiste italien Vilfredo Pareto (1848-1923), est aussi connue sous le nom de loi de Bradford. Cette loi a d'abord été utilisée pour représenter la répartition des richesses dans la société, avec notamment le principe de Pareto, selon lequel 80% des richesses d'un pays sont détenus par 20% de la population.

- PERT (a, m, b) : la densité de cette loi est donnée par :

$$f(x) = \frac{1}{B(\alpha,\beta)} \frac{(x-a)^{\alpha-1}(b-x)^{\beta-1}}{(b-a)^{\alpha+\beta-1}}, \text{ avec } \alpha,\beta > 0, x \in [a,b]$$

$$a,b \in \mathbb{R} \text{ et } B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

$$\alpha = \frac{4m+b-5a}{b-a}$$

$$\beta = \frac{5b-a-4m}{b-a}$$

On a $E(X) = (b-a)\alpha/(\alpha+\beta)$ et $V(X) = (b-a)^2\alpha\beta/\left((\alpha+\beta+1)\left(\alpha+\beta\right)^2\right)$

La loi de PERT est donc un cas particulier de la loi Bêta4, définie par son intervalle de définition $[a,b]$ et sa valeur la plus probable m (le mode). PERT est l'acronyme de *Program Evaluation and Review Technique* , une méthode de gestion et de planification de projet. La méthodologie et la distribution PERT ont été utilisées pour la première fois pour le projet de développement des missiles Polaris lancés depuis des sous-marins par la marine américaine et Lockheed de 1956 à 1960 (Clark 1962). La distribution PERT permet de modéliser le temps probable nécessaire à une équipe pour terminer son projet. La loi triangulaire, plus simple, permet aussi de modéliser ce type de phénomènes avec les trois mêmes paramètres.

- Poisson (λ): la densité de cette loi est donnée par :

$$P(X=x) = \frac{\exp(-\lambda)\lambda^x}{x!}, \quad \text{avec } x \in \mathbb{N} \text{ et } \lambda > 0$$

On a $E(X) = \lambda$ et $V(X) = \lambda$

La loi de Poisson, découverte par le mathématicien et astronome Siméon-Denis Poisson (1781-1840) qui fut élève de Laplace, Lagrange et Legendre, est souvent utilisée pour étudier des phénomènes de file d'attente.

- Student (df) : la densité de cette loi est donnée par :

$$f(x) = \frac{\Gamma((df+1/2))}{\sqrt{\pi df}\Gamma(df/2)}(1+x^2/df)^{-(df+1)/2}, \text{ avec } df > 0$$

On a $E(X) = 0$ si $df > 1$ et $V(X) = df/(df-2)$ si $df > 2$

La loi de Student, du nom que se donnait le chimiste et statisticien anglais William Sealy Gosset (1876-1937) afin de préserver son anonymat (la brasserie Guinness interdisait à ses employés de publier, suite à la publication par un autre chercheur d'informations confidentielles) est la loi de la moyenne de df variables distribuées suivant une loi normale centrée réduite. Lorsque $df = 1$, la loi de Student est une loi de Cauchy dont la particularité est de n'avoir ni espérance ni variance.

- Trapézoïdale (a, b, c, d) : la densité de cette loi est donnée par :

$$\begin{cases} f(x) = \frac{2(x-a)}{(d+c-b-a)(b-a)}, & x \in [a, b] \\ f(x) = \frac{2}{(d+c-b-a)}, & x \in [b, c] \\ f(x) = \frac{2(d-x)}{(d+c-b-a)(d-c)}, & x \in [c, d] \\ f(x) = 0, & x < a, \quad x > d \\ \text{avec } a < b < c < d \end{cases}$$

On a $E(X) = \frac{d^2+c^2-b^2-a^2+cd-ab}{3(d+c-b-a)}$ et $V(X) = \frac{(c+d)(c^2+d^2)-(a+b)(a^2+b^2)}{6(d+c-b-a)} - E^2(X)$

Cette loi est utile pour représenter un phénomène dont on sait qu'il peut prendre des valeurs entre deux extrêmes, mais pour lequel un intervalle plus restreint paraît plus raisonnable.

- Triangulaire (a, m, b) : la densité de cette loi est donnée par :

$$\begin{cases} f(x) = \frac{2(x-a)}{(b-a)(m-a)}, & x \in [a, m] \\ f(x) = \frac{2(b-x)}{(b-a)(b-m)}, & x \in [m, b] \\ f(x) = 0, & x < a, \quad x < b \\ \text{avec } a < m < b \end{cases}$$

On a $E(X) = (a + m + b)/3$ et $V(X) = (a^2 + m^2 + b^2 - ab - am - bm)/18$

- TriangulaireQ (q_1, m, q_2, p_1, p_2): cette loi est une reparamétrisation de la loi triangulaire. Une première étape nécessite l'estimation des paramètres a et b de la distribution triangulaire pour savoir à quels quantiles q_1 et q_2 correspondent les pourcentages p_1 et p_2. Une fois ceci fait, on peut utiliser la fonction de densité ou de répartition triangulaire.

- Uniforme (a, b) : la densité de cette loi est donnée par :

$$f(x) = \frac{1}{b-a}, \text{ avec } b > a \text{ et } x \in [a, b]$$

On a $E(X) = (a + b)/2$ et $V(X) = (b - a)^2/12$

La loi uniforme (0, 1) est très utilisée pour les simulations. Comme la fonction de répartition de toutes les lois est comprise entre 0 et 1, un échantillon tiré dans une loi Uniforme (0,1) permet

d'obtenir un échantillon dans toutes les lois dont on sait calculer l'inverse.

- Uniforme discrète (a, b) : la densité de cette loi est donnée par :

$$P[X = x] = \frac{1}{b - a + 1}, \text{ avec } (a, b, x) \in \mathbb{N}^3, x \in [a, b]$$

On a $E(X) = (a + b)/2$ et $V(X) = [(b - a + 1)^2 - 1]/12$

La loi uniforme discrète correspond au cas particulier où la loi uniforme est restreinte à des nombre entiers.

- Weibull (β) : la densité de cette loi est donnée par :

$$f(x) = \beta x^{\beta-1} \exp(-x^\beta), \text{ avec } x > 0 \text{ et } \beta > 0$$

On a $E(X) = \Gamma(\frac{1}{\beta} + 1)$ et $V(X) = \Gamma(\frac{2}{\beta} + 1) - \Gamma^2(\frac{1}{\beta} + 1)$

Le paramètre β est le paramètre de forme de la loi de Weibull.

- Weibull (β, γ) : la densité de cette loi est donnée par :

$$f(x) = \frac{\beta}{\gamma}(\frac{x}{\gamma})^{\beta-1} e^{-(\frac{x}{\gamma})^\beta}, \text{ avec } x > 0, \text{ et } \beta, \gamma > 0$$

On a $E(X) = \gamma\Gamma\left(\frac{1}{\beta} + 1\right)$ et $V(X) = \gamma^2\left[\Gamma\left(\frac{2}{\beta} + 1\right) - \Gamma^2\left(\frac{1}{\beta} + 1\right)\right]$

Le paramètre β est le paramètre de forme et le paramètre γ est le paramètre d'échelle. Lorsque $\beta = 1$, la loi de Weibull est une loi exponentielle de paramètre $1/\gamma$.

- Weibull (β, γ, μ): la densité de cette loi est donnée par :

$$f(x) = \frac{\beta}{\gamma}\left(\frac{x - \mu}{\gamma}\right)^{\beta-1} e^{-\left(\frac{x-\mu}{\gamma}\right)^\beta}, \text{ avec } x > \mu, \text{ et } \beta, \gamma > 0$$

On a $E(X) = \mu + \gamma\Gamma\left(\frac{1}{\beta} + 1\right)$ et $V(X) = \gamma^2\left[\Gamma\left(\frac{2}{\beta} + 1\right) - \Gamma^2\left(\frac{1}{\beta} + 1\right)\right]$

La loi de Weibull, du nom du suédois Ernst Hjalmar Waloddi Weibull (1887-1979), est très utilisée en contrôle qualité et en analyse de survie. Le paramètre β est le paramètre de forme et le paramètre γ est le paramètre d'échelle. Lorsque $\beta = 1$ et $\mu = 0$, la loi de Weibull est une loi exponentielle de paramètre $1/\gamma$.

Boîte de dialogue

OK : cliquez sur ce bouton pour lancer les calculs.

Annuler : cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.

Distribution théorique : activez cette option pour échantillonner des données dans une loi de distribution théorique. Veuillez alors choisir la loi, puis entrez les paramètres de la loi si nécessaire.

Distribution empirique : activez cette option pour échantillonner des données dans une loi empirique. Sélectionnez alors les données permettant de construire la loi empirique.

Libellés des colonnes : activez cette option si la première ligne des données sélectionnées (données et éventuellement poids) contient un libellé.

Poids : activez cette option si vous voulez pondérer l'échantillonnage. Les poids doivent être impérativement supérieurs ou égaux à 0. Si un en-tête de colonne a été sélectionné, veuillez vérifier que l'option « Libellés des colonnes » est activée.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Nombre d'échantillons : entrez le nombre de colonnes à générer.

Taille d'échantillon : entrez le nombre de données à générer pour chacun des échantillons.

Afficher l'en-tête du rapport : désactivez cette option si vous souhaitez que le tableau des données échantillonnées commence dès la première ligne de la feuille Excel (cas d'une sortie dans une feuille ou un classeur), et non après l'en-tête du rapport.

Exemple

Un exemple de génération d'un échantillon aléatoire tiré dans une loi normale est disponible sur le Centre d'aide XLSTAT :

http://www.xlstat.com/demo-normf.htm

Bibliographie

Abramowitz M. and Stegun I.A. (1972). Handbook of Mathematical Functions. Dover Publications, New York.

Clark C. E. (1962). The PERT model for the distribution of an activity time. *Operation Research*, **10** (3), 405-406.

El-Shaarawi A.H., Esterby E.S. and Dutka B.J (1981). Bacterial density in water determined by Poisson or negative binomial distributions. *Applied an Environmental Microbiology*, **41** (1). 107-116.

Fisher R.A. and Tippett H.C. (1928). Limiting forms of the frequency distribution of the smallest and largest member of a sample. *Proc. Cambridge Phil. Soc.*, **24**, 180-190.

Gumbel E.J. (1941). Probability interpretation of the observed return periods of floods. *Trans. Am. Geophys. Union*, **21**, 836-850.

Jenkinson A. F. (1955). The frequency distribution of the annual maximum (or minimum) of meteorological elements. *Q. J. R. Meteorol. Soc.*, **81**, 158-171.

Perreault L. and Bobée B. (1992). Loi généralisée des valeurs extrêmes. Propriétés mathématiques et statistiques. Estimation des paramètres et des quantiles XT de période de retour T. INRS-Eau, rapport de recherche no 350, Québec.

Weibull W. (1939). A statistical theory of the strength of material. *Proc. Roy. Swedish Inst. Eng. Res.* **151** (1), 1-45.

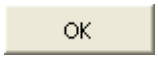
Transformation de variables

Utilisez cet outil pour transformer rapidement une ou plusieurs variables.

Dans cette section :

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.


: cliquez sur ce bouton pour changer la façon dont vous chargez les données dans XLSTAT. Si le bouton possède une icône de souris, XLSTAT vous permet de faire une sélection à la souris des données. Si le bouton possède une icône de liste, XLSTAT vous permet de sélectionner les données à partir d'une liste. Si le bouton possède une icône orange, des boutons complémentaires avec un point d'interrogation  sont affichés pour vous permettre d'importer les données à partir de fichiers.

Onglet **Général** :

Données : sélectionnez les données sur la feuille Excel. Si des en-têtes ont été sélectionnés, veuillez vérifier que l'option « Libellés des colonnes » est activée.

Transformation :

- **Normaliser (n-1)** : choisissez cette option pour normaliser les variables en utilisant l'écart-type non biaisé.
- **Autre** : choisissez cette option pour utiliser une autre transformation. Cliquez alors sur l'onglet transformation pour choisir une autre transformation.
    - **Normaliser (n)** : choisissez cette option pour normaliser les variables en utilisant l'écart-type biaisé.
    - **Centrer** : choisissez cette option pour centrer les données.
    - **/ Ecart-type (n-1)** : choisissez cette option pour diviser les données par l'écart-type non biaisé.
    - **/ Ecart-type (n)** : choisissez cette option pour diviser les données par l'écart-type biaisé.
    - **Remettre à l'échelle de 0 à 1** : choisissez cette option pour transformer les données de telle sorte qu'elles soient comprises entre 0 et 1.
    - **Remettre à l'échelle de 0 à 100** : choisissez cette option pour transformer les données de telle sorte qu'elles soient comprises entre 0 et 100.
    - **Pareto scaling** : choisissez cette option pour normaliser les variables en utilisant la racine carrée de l'écart-type.
    - **Binariser (0/1)** : choisissez cette option pour transformer les données de telle sorte que les données égales à 0 soient égales à 0, et les données différentes de 0 soient égales à 1.
    - **Signe (-1/0/1)** : choisissez cette option pour transformer les données de telle sorte que les données égales à 0 soient égales à 0, les données négatives soient égales à -1 et les données positives soient égales à 1.
    - **Arc sinus** : choisissez cette option pour calculer l'arc-sinus des données sélectionnées.
    - **Transformation Box-Cox** : activez cette option pour faire une transformation de Box-Cox. Vous pouvez soit imposer une valeur de **Lambda**, soit décider que XLSTAT doit l'**optimiser**. Cette transformation permet d'augmenter la normalité des données ; l'équation de Box-Cox est définie par : $$ Y_{t} = \left{\begin{array} { l l } {\frac { X _ { t } ^ { \lambda } - 1} { \lambda } ,} & { (X _ { t } > 0,\lambda \neq 0) \text{ ou } (X _ { t } \geq 0,\lambda > 0) } \ { \ln (X _ { t }) ,} & { X _ { t } > 0,\lambda = 0} \end{array} \right. $$ Si l'option d'optimisation est choisie, XLSTAT maximise la vraisemblance de l'échantillon, étant supposé qu'après transformation l'échantillon suit une loi normale.

- **Winsorize** : utilisez cette transformation pour éliminer les données ne correspondant pas à un intervalle donné par deux percentiles : soit **p1** et **p2** deux valeurs comprises entre 0 et 1, telles que $p1 < p2$. Si une valeur x de l'échantillon est inférieure à $q1$, le quantile correspondant à $p1$ obtenu à partir de l'échantillon, ou supérieure à $q2$ le quantile correspondant à $p2$, alors la valeur est transformée en $q1$ dans le premier cas et en $q2$ dans le second cas.

- **Transformation de Johnson** : choisissez la transformation Johnson pour transformer vos données afin de suivre une distribution normale. Cette transformation est une généralisation de la transformation Box-Cox qui ne s'applique qu'aux valeurs positives. La sélection de la distribution et l'estimation des paramètres sont effectuées en utilisant l'approche décrite par Chou *et al.* (1998) :

- La famille S_B de Johnson : $$ Y_{t} = \gamma + \eta ln(\frac{X_t - \epsilon}{\lambda + \epsilon - X_t}) $$ avec $\eta, \lambda > 0, -\infty < \gamma < \infty, -\infty < \epsilon < \infty$ et $\epsilon < X_t < \epsilon + \lambda$

- La famille S_L de Johnson : $$ Y_{t} = \gamma + \eta ln(X_t - \epsilon) $$ avec $\eta > 0, -\infty < \gamma < \infty, -\infty < \epsilon < \infty$ et $\epsilon < X_t$

- La famille S_U de Johnson : $$ Y_{t} = \gamma + \eta sinh^{-1}(\frac{X_t - \epsilon}{\lambda}) $$ avec $\eta, \lambda > 0, -\infty < \gamma < \infty, -\infty < \epsilon < \infty$ et $-\infty < X_t < \infty$

Vous pouvez choisir le test de normalité utilisé pour identifier la meilleure transformation ainsi que le niveau de signification.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des colonnes : activez cette option si la première ligne des données sélectionnées (données et tableau de codage) contient un libellé.

Afficher l'en-tête du rapport : désactivez cette option si vous souhaitez que le tableau des résultats commence dès la première ligne de la feuille Excel (cas d'une sortie dans une feuille ou un classeur), et non après l'en-tête du rapport.

Libellés des observations : activez cette option si vous voulez utiliser des libellés d'observations disponibles sur une feuille Excel pour l'affichage des résultats. Si l'option « Libellés des colonnes » est activée, la première cellule de la sélection doit comprendre un en-tête. Si vous n'activez pas cette option, des libellés seront automatiquement créés (Obs1, Obs2, …).

Onglet **Données manquantes** :

Ne pas accepter les valeurs manquantes : activez cette option pour que XLSTAT empêche la poursuite des calculs si des valeurs manquantes sont détectées.

Supprimer les observations : activez cette option pour supprimer les observations comportant des données manquantes.

Ignorer les données manquantes : activez cette option pour ignorer les données manquantes.

Estimer les données manquantes : activez cette option pour estimer les données manquantes avant le début des calculs.

- **Moyenne ou mode** : activez cette option pour estimer les données manquantes en utilisant la moyenne (variables quantitatives) ou le mode (variables qualitatives) pour les variables correspondantes.
- **Plus proche voisin** : activez cette option pour estimer les données manquantes d'une observation en recherchant le plus proche voisin de l'observation.

Onglet **Sorties** :

Formule générale : activez cette option pour afficher la formule générale qui est utilisée pour la transformation choisie (excepté pour la transformation de Johnson).

Formule par variable : activez cette option pour afficher la formule exacte utilisée par la transformation choisie (excepté pour les transformations "Binariser", "Signe" et "Arc sinus").

Exemple

Un exemple d'utilisation de la transformation de Johnson est disponible sur le Centre d'aide XLSTAT à l'adresse

http://www.xlstat.com/demo-johnsonFR.htm

Un exemple d'utilisation de la transformation Box-Cox est disponible sur le Centre d'aide XLSTAT à l'adresse

https://help.xlstat.com/fr/6640-transformation-box-cox-dans-excel

Bibliographie

Chou Y-M, Polansky A. M. and Mason R.L. (1998). Transforming Non-Normal Data to Normality in Statistical Process Control. *Journal of Quality Technology*, **30:2**, 133-141

Johnson N. L. (1949). Systems of frequency curves generated by methods of translation. *Biometrika*, **36**, 149-176.

Anonymisation des données

Utilisez cet outil pour transformer vos données en données anonymisées. Trois méthodes sont proposées : séquentielle, aléatoire et tableau des correspondances. Ces méthodes ont l'avantage de s'appliquer aussi bien sur des données quantitatives que qualitatives.

Dans cette section :

Description

Pour partager des données privées parfois sensibles il est intéressant de pouvoir les transformer. XLSTAT propose trois méthodes de transformation :

Séquentielle

Les modalités de toutes les variables sélectionnées sont remplacées par un chiffre entier choisi de façon séquentielle en commençant à 1. Le chiffre associé à une modalité apparait autant de fois que la modalité apparaît dans le jeu de données. Le fichier résultat est ainsi un fichier numérique.

Aléatoire

La méthode varie selon le type de la variable. Pour une variable quantitative, ses valeurs sont mélangées aléatoirement entre elles. Ainsi, elles restent dans le même ordre de grandeur que les données initiales. Pour les variables qualitatives les modalités sont remplacées par une chaine de caractères choisis aléatoirement. Cette chaîne est utilisée autant de fois que la modalité apparaît dans le jeu de données.

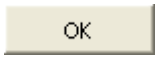
Tableau des correspondances

Les données sont remplacées par les nouvelles valeurs fournies par le tableau des correspondances. Ce dernier précise la valeur originale des variables à remplacer dans la colonne de gauche et les nouvelles valeurs dans la colonne de droite.

Pour compléter l'outil, une option d'anonymisation des noms de variables est aussi proposée.

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT vous permet de sélectionner les données par colonnes ou par plage. Si la flèche est vers la droite, XLSTAT vous permet de sélectionner les données par lignes ou par plage.


: cliquez sur ce bouton pour changer la façon dont vous chargez les données dans XLSTAT. Si le bouton possède une icône de souris, XLSTAT vous permet de faire une sélection à la souris des données. Si le bouton possède une icône de liste, XLSTAT vous permet de sélectionner les données à partir d'une liste. Si le bouton possède une icône orange, des boutons complémentaires avec un point d'interrogation  sont affichés pour vous permettre d'importer les données à partir de fichiers.

Onglet **Général** :

Données : sélectionnez les données que vous souhaitez transformer. Si des en-têtes de colonnes ont été sélectionnés, veuillez vérifier que l'option « Libellés des variables » est activée.

Variables anonymisées : activez cette option si vous souhaitez que le nom de toutes les variables soit aussi anonymisé.

Plage : activez cette option pour que les résultats soient affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des variables : activez cette option si la première ligne des données sélectionnées contient un libellé.

Libellés des observations : activez cette option pour sélectionner les libellés d'observations qui seront ensuite utilisés pour l'affichage des résultats. Si l'option « Libellés des variables » est activée, la première cellule de la sélection doit comprendre un en-tête. Si vous n'activez pas cette option, des libellés seront automatiquement créés (Obs1, Obs2, …).

Onglet **Options** :

Méthodes d'anonymisation : sélectionnez la méthode d'anonymisation parmi les 3 proposées : séquentielle, aléatoire, tableau des correspondances.

Supprimer les espaces : activez cette option pour que XLSTAT supprime les espaces dans chaque cellule sélectionnée avant (cochez à gauche) et/ou après (cochez à droite) des valeurs.

Onglet **Données manquantes** :

Ne pas accepter les valeurs manquantes : activez cette option pour que XLSTAT empêche la poursuite des calculs si des valeurs manquantes sont détectées.

Supprimer les observations : activez cette option pour supprimer les observations comportant des données manquantes.

Onglet **Sorties** :

Données d'origine : activez cette option pour afficher le tableau des données qui ont été sélectionnées.

Données anonymisées : activez cette option pour afficher le tableau des données anonymisées.

Tableau de correspondance : activez cette option pour afficher le tableau de correspondance entre les données originales et les données transformées.

Résultats

Données d'origine : ce tableau regroupe l'ensemble des données sélectionnées tel qu'il est affiché dans la feuille de données.

Données anonymisées : ce tableau regroupe l'ensemble des données qui ont été anonymisées. Elles sont affichées dans le même ordre que les données initiales. Si l'option "Variables anonymisées" a été choisie, alors les libellés de ces variables sont présentés dans leur forme anonyme.

Tableau de correspondance : ce tableau liste les modalités des variables qui ont été transformées. Dans la colonne de gauche sont regroupées les valeurs initiales et dans la colonne de droite la nouvelle valeur.

Exemple

Un tutoriel sur l'utilisation du module XLSTAT-Anonymisation des données est disponible sur le Centre d'aide XLSTAT :

http://www.xlstat.com/demo-cryf.htm

Données manquantes

Utilisez cet outil pour traiter un jeu de données avec des données manquantes préalablement à d'autres analyses avec XLSTAT.

Dans cette section :

Description

La plupart des outils d'XLSTAT comportent un onglet pour le traitement des données manquantes. Néanmoins, les méthodes disponibles sont peu nombreuses. Cet outil vous permet de prétraiter vos données en complétant les données manquantes avec des méthodes avancées.

Il existe trois types de données manquantes (Allison, 2001) : les données manquantes de façon complètement aléatoire (MCAR), les données manquantes aléatoirement (MAR) et les données manquantes de façon non aléatoire (NMAR).

Si ce qui a amené à ce qu'une donnée soit manquante ne dépend d'aucune variable observable et d'aucun paramètre non observable, alors les données manquent de façon complètement aléatoirement (MCAR). Le fait qu'une donnée manque est alors considéré comme dû au hasard. Dans ce cas, les analyses effectuées sont non biaisées.

Si ce qui a amené à ce qu'une donnée soit manquante est lié à la valeur d'une variable externe mais pas aux valeurs de la variable ayant des données manquantes, alors les données manquent de façon aléatoire (MAR). C'est le cas le plus classique.

Si les données manquent pour une raison particulière, alors les données manquent non aléatoirement (NMAR). Un exemple classique est le cas des questions filtrées (certaines questions ne concernent que certaines personnes dans un questionnaire, les autres personnes sont manquantes).

Les méthodes disponibles dans cet outil permettent de traiter les cas MCAR et MAR.

Différentes méthodes sont disponibles en fonction du type de données et de vos besoins :

- Pour des données quantitatives, XLSTAT vous permet de :
- Supprimer les observations ayant des données manquantes.

- Utiliser une imputation par la moyenne de chaque variable.
- Utiliser une approche de plus proche voisin.
- Remplacer les valeurs manquantes par une valeur numérique donnée.
- Utiliser l'algorithme NIPALS.
- Utiliser une méthode d'imputation multiple utilisant les MCMC (Markov Chain Monte Carlo).
- Utiliser l'algorithme EM (Expectation Maximisation) pour des données suivant une loi normale multivariée.
- Pour des données qualitatives, XLSTAT vous permet de :
- Supprimer les observations ayant des données manquantes.
- Utiliser une imputation par le mode de chaque variable.
- Utiliser une méthode de plus proche voisin.
- Remplacer les valeurs manquantes par une valeur textuelle donnée.
- Utiliser l'algorithme NIPALS.

Algorithme NIPALS

L'algorithme NIPALS est une méthode présentée par H. Wold (1973) permettant d'effectuer une analyse en composantes principales sur les données disponibles. L'algorithme NIPALS est appliqué sur les données pour obtenir un modèle d'ACP. Ce modèle est ensuite utilisé pour prédire les données manquantes.

Imputation multiple

XLSTAT propose un algorithme d'imputation multiple basé sur les chaines de Markov (MCMC), il est aussi appelé *fully conditional specification* (Van Buulen, 2007).

L'algorithme fonctionne de la manière suivante :

1. Des valeurs initiales pour les données manquantes sont obtenues en tirant aléatoirement des valeurs sur une loi normale de moyenne et variance égale à la moyenne et à la variance obtenues sur les données disponibles.
2. Pour chaque variable du jeu de données ayant des données manquantes, une méthode d'imputation basée sur l'échantillonnage dans une distribution et le modèle MCO est appliquée. Le modèle utilisé est un modèle de régression ayant la variable étudiée comme variable dépendante et les autres variables du jeu de données comme variables indépendantes. Des valeurs aléatoires tirées sur des lois définies sont utilisées pour

apporter une part aléatoire au modèle. Les valeurs imputées sont obtenues à partir du modèle estimé.

Ces deux étapes sont répétées autant de fois que demandé par l'utilisateur. La valeur moyenne de chaque donnée manquante imputée est utilisée.

Algorithme EM

L'algorithme EM utilisé par XLSTAT doit être appliqué sur des données suivant une **loi normale multivariée**. A noter que les poids des observations ne sont pas pris en compte avec cette méthode.

L'algorithme EM (Dempster, Laird et Rubin, 1977) est une technique permettant d'estimer le maximum de vraisemblance (MLE) pour des données incomplètes. Pour une description détaillée sur les applications de l'algorithme EM, voir les livres de Little et Rubin (2002) et Schafer (1997).

L'algorithme EM est une procédure itérative qui trouve le MLE des paramètres de la loi normale multivariée en répétant les étapes suivantes :

1. L'étape E (Expectation) : étant donné un ensemble de paramètres (vecteur de moyennes et matrice de variance covariance pour une distribution normale multivariée), l'étape E calcule la vraisemblance attendue des données complètes compte tenu des données observées et des estimations des paramètres.

2. L'étape M (Maximisation) : à partir de la vraisemblance des données complètes, l'étape M trouve les estimations des paramètres maximisant la vraisemblance des données complètes obtenue à l'étape E.

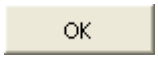
Les deux étapes sont répétées jusqu'à convergence.

Remarques

Lorsque vous disposez de variables quantitatives et de variables qualitatives, et que vous choisissez d'estimer les valeurs manquantes, alors celles ci sont traitées de manière indépendantes sur les deux tableaux de données. Par contre si vous choisissez de supprimer les lignes avec des valeurs manquantes, alors les deux tableaux seront fusionnés et les lignes supprimées sur un tableau seront également supprimées sur l'autre tableau.

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.


: cliquez sur ce bouton pour changer la façon dont vous chargez les données dans XLSTAT. Si le bouton possède une icône de souris, XLSTAT vous permet de faire une sélection à la souris des données. Si le bouton possède une icône de liste, XLSTAT vous permet de sélectionner les données à partir d'une liste. Si le bouton possède une icône orange, des boutons complémentaires avec un point d'interrogation  sont affichés pour vous permettre d'importer les données à partir de fichiers.

Onglet **Général**:

Données quantitatives : activez cette option pour sélectionner le tableau de variables quantitatives contenant des données manquantes. Si des en-têtes de colonnes ont été sélectionnés, veuillez vérifier que l'option « Libellés des variables » est activée.

Supprimer les observations : activez cette option pour supprimer les lignes comportant des données manquantes.

Estimer les données manquantes : activez cette option pour estimer les données manquantes avec la méthode de votre choix (voir la section Description pour plus de détails sur les méthodes disponibles).

- **Valeur** : si vous avez choisi d'estimer toutes les données manquantes avec la même valeur, entrez la valeur numérique à utiliser.
- **Nombre d'imputations** : si vous avez sélectionné la méthode d'imputation multiple, entrez le nombre d'imputations à effectuer.
- **Itérations** : si vous avez sélectionné l'algorithme EM, entrez le nombre d'itérations à effectuer.
- **Convergence** : si vous avez sélectionné l'algorithme EM, entrez le niveau de convergence souhaité.

Données qualitatives : activez cette option pour sélectionner le tableau de variables qualitatives contenant des données manquantes. Si des en-têtes de colonnes ont été sélectionnés, veuillez vérifier que l'option « Libellés des variables » est activée.

Supprimer les observations : activez cette option pour supprimer les lignes comportant des données manquantes.

Estimer les données manquantes : activez cette option pour estimer les données manquantes avec la méthode de votre choix (voir la section Description pour plus de détails sur les méthodes disponibles).

- **Valeur** : si vous avez choisi d'estimer toutes les données manquantes avec la même valeur, entrez la valeur textuelle à utiliser.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des variables : activez cette option si la première ligne des données sélectionnées (données, libellés des observations, poids) contient un libellé.

Libellés des observations : activez cette option si vous voulez utiliser des libellés d'observations pour l'affichage des résultats. Si l'option « Libellés des variables » est activée, la première cellule de la sélection doit comprendre un en-tête. Si vous n'activez pas cette option, des libellés seront automatiquement créés (Obs1, Obs2, …).

Poids : activez cette option si vous voulez pondérer les observations. Si vous n'activez pas cette option, les poids seront tous considérés comme valant 1. Les poids doivent être impérativement supérieurs ou égaux à 0. Si un en-tête de colonne a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

Onglet **Sorties**:

Statistiques descriptives : activez cette option pour afficher les statistiques descriptives associées aux données avant et après imputation.

Graphique des données manquantes : activez cette option pour afficher le graphique des données manquantes. Ce graphique représente votre jeu de données avec les données manquantes colorées en rouge. Les libellés des variables contiennent également le pourcentage de données manquantes par colonne.

Résultats de l'ACM : activez cette option pour afficher le tableau des coordonnées principales obtenu par l'analyse des correspondances multiples (ACM) sur les données manquantes. Pour chaque variable, la modalité '0' représente la donnée présente alors que la modalité '1' modélise les données manquantes. La carte graphique de ce résultat s'affiche automatiquement en dessous.

Résultats

Si vous avez coché l'ensemble des sorties proposées alors pour chaque type de données (quantitatives ou qualitatives) vous aurez dans cet ordre : - le graphique des données manquantes, - les tableaux des statistiques descriptives avant traitement et après traitement, - le tableau des données complétées dont les valeurs imputées sont affichées en gras, - le tableau et le graphique de l'ACM sur les données manquantes pour chaque variable.

Exemple

Des exemples de traitement des données manquantes sont disponibles à l'adresse suivante :

https://www.xlstat.com/demo-missingf.htm https://www.xlstat.com/demo-missingf2.htm

Bibliographie

Allison P. D. (Ed.). (2001). Missing data (No. 136). Sage.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, **39(1), 1-22**.

Josse J. (2016) Contribution to missing values & principal component methods. HDR Statistics. Université Paris Sud - Orsay, 2016.

Schafer J. L. (1997). Analysis of Incomplete Multivariate Data. London: Chapman and Hall.

Van Buuren S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research*, **16**, 219–242.

Wold H. (1973). Non-linear Iterative PArtial Least Squares (NIPALS) modelling. Some current developments. In: P.R. Krishnaiah (Ed.), Multivariate Analysis III, Academic Press, New York, 383-407.

Redressement de sondage

Utilisez cet outil pour calculer les poids afin de redresser un sondage en utilisant des variables auxiliaires dont on connaît les proportions sur la population.

Dans cette section :

Description

Dans la théorie des sondages, le redressement constitue un point très important. Il arrive souvent que l'on possède un certain nombre de variables qualitatives, dites auxiliaires, dont on connaît la répartition sur l'ensemble de la population étudiée. On veut alors se servir de ces variables afin de redresser un sondage et ainsi d'obtenir des répartitions comparables pour la population et l'échantillon. La recherche de ces poids de sondage utilise des méthodes statistiques avancées.

XLSTAT vous propose quatre méthodes de calcul des poids de redressement dont la méthode raking ratio (Deming et Stefan, 1940).

Soit un échantillon de taille n sur une population dont on connaît la taille N. On effectue un sondage sur cet échantillon en incluant dans ce sondage un certain nombre de variables auxiliaires dont on connaît la répartition sur la population globale. Les méthodes de redressements vont permettre par des algorithmes itératifs de trouver les poids adaptés afin que l'échantillon « ressemble » le plus possible à la population.

Les 4 méthodes présentent dans XLSTAT sont :

- Méthode raking ratio
- Méthode logit : celle-ci correspond à la méthode raking ratio en fixant des contraintes de bornes sur les poids
- Méthode linéaire
- Méthode linéaire tronquée : celle-ci correspond à la méthode linéaire avec des bornes sur les poids

Ces méthodes sont basées sur le même algorithme développé par Deville, Särndall et Sautory (1993) et doivent aboutir à des résultats proches. La principale différence réside dans la fonction à optimiser.

Méthodes de calcul des poids de redressement :

XLSTAT permet d'utiliser quatre méthodes afin de redresser un échantillon. Elles sont toutes calculées en utilisant la méthode appelée « generalized raking procedure ». Le cas raking ratio est un cas particulier de cette méthode.

Soit un échantillon composé de n observations et p variables qualitatives dites auxiliaires, on notera x_{ij} les éléments de cet échantillon. D'autre part, on possède la répartition des p variables sous forme de sommes marginales sur la population, notées X_j. Soit d_i les poids associés aux observations avant le redressement et w_i, les poids associés aux observations après le redressement.

On veut donc trouver des poids w_i proches des d_i qui satisfont les équations de calage suivantes :

$$\sum_{k=1}^{n} w_k x_{kj} = X_j \quad \forall j = 1, \ldots, p$$

On choisit donc une fonction de distance $G()$ d'argument w_k/d_k, cette fonction doit être convexe et positive. On a donc un problème d'optimisation sous contraintes qui peut être résolu en utilisant la méthode des multiplicateurs de Lagrange (λ).

Le problème est le suivant :

$$\min_{w_k} \sum_{k=1}^{n} d_k G(W_k/d_k)$$
$$s.c. \sum_{k=1}^{n} w_k x_k = X$$

Le Lagrangien vaut :

$$L = \sum_{k=1}^{n} d_k G(W_k/d_k) - \lambda'(\sum_{k=1}^{n} w_k x_k - X)$$

Nous avons donc :

$$w_k = d_k F(x_k'\lambda)$$
$$\sum_{k=1}^{n} d_k F(x_k'\lambda) x_k = X$$

Avec $F()$ fonction réciproque de la dérivée de $G()$.Ce système d'équations peut être résolu en utilisant une méthode tel que la méthode de Newton.

Les fonctions $F()$ sont les suivantes :

- Pour la méthode raking ratio nous avons :

$$F(u) = \exp(u)$$

- Pour la méthode logit, nous avons pour des bornes L (inférieure) et U (supérieure) :

$$F(u) = \frac{L(U-1) + U(1-L)\exp(Au)}{U - 1 + (1-L)\exp(Au)}$$

$$A = \frac{U-L}{(1-L)(U-1)}$$

- Pour la méthode linéaire, nous avons ;

$$F(u) = 1 + u$$

- Pour la méthode linéaire tronquée, nous avons :

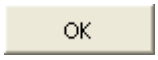
$$F(u) = 1 + u \in [L; U]$$

L'algorithme s'arrête lorsque le critère suivant atteint une valeur epsilon définie par l'utilisateur :

$$\max_{k=1}^{n} \left| \frac{w_k^{(i+1)}}{d_k} - \frac{w_k^{(i)}}{d_k} \right| < \epsilon$$

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.

Onglet **Général**:

Données de redressement : sélectionner les échantillons de données auxiliaires qualitatives. Si des en-têtes ont été sélectionnés, veuillez vérifier que l'option « Libellés des échantillons » est activée.

Sommes marginales sur la population : sélectionner les sommes marginales pour les variables auxiliaires sur la population. Les variables doivent être représentées par colonne avec une ligne par modalité de la variable. Les colonnes doivent être sélectionnées dans le même ordre que les données de redressement.

Format :

- **Valeurs :** activez cette option si les valeurs des sommes marginales sur la population représentent la valeur réelle dans la population.
- **Pourcentages :** activez cette option si les valeurs des sommes marginales sur la population sont sous forme de pourcentages. Dans ce cas, il vous faut spécifier la **taille de la population**.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des échantillons : activez cette option si la première ligne des sélections (données qualitatives, poids) contient un libellé.

Poids initiaux : activez cette option si vous voulez utiliser une pondération initiale. Si vous n'activez pas cette option, les poids seront tous considérés comme valant N/n (taille de la population / taille de l'échantillon). Les poids doivent être impérativement supérieurs à 0. Si un en-tête de colonne a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

Libellés des observations : activez cette option si vous voulez utiliser des libellés d'observations disponibles sur une feuille Excel pour l'affichage des résultats. Si l'option « Libellés des variables » est activée, la première cellule de la sélection doit comprendre un en-tête. Si vous n'activez pas cette option, des libellés seront automatiquement créés (Obs1, Obs2, …).

Onglet **Options**:

Méthode d'estimation : sélectionner la méthode d'estimation utilisée. Pour les méthodes logit et linéaire tronqué, les bornes inférieure et supérieure doivent être renseignées. On doit avoir

borne inférieure < 1 et borne supérieure > 1.

Conditions d'arrêt :

- **Itérations** : entrez le nombre maximal d'itérations pour l'algorithme. Les calculs sont interrompus dès que le nombre maximal d'itérations est dépassé. Valeur par défaut : 50.
- **Convergence** : entrez la valeur seuil du critère de convergence de l'algorithme. Valeur par défaut : 0,0001.

Onglet **Données manquantes** :

Supprimer les observations : activez cette option pour ne pas prendre en compte une observation dont l'une des données est manquante.

Estimer les données manquantes : activez cette option pour estimer les données manquantes en utilisant le mode de l'échantillon.

Onglet **Sorties**:

Statistiques descriptives : activez cette option pour afficher les statistiques descriptives des échantillons avant et après le redressement.

Données en sortie : activez cette option pour afficher dans le tableau des poids les variables auxiliaires originales.

Rapport de poids : activez cette option pour afficher dans le tableau des poids les rapports de poids (poids final / poids initial)

Liste des combinaisons : activez cette option pour afficher sous forme de tableau l'ensemble des combinaisons des modalités des variables auxiliaires avec leur effectif et le rapport de poids associé.

Détails des itérations : activez cette option pour afficher les détails des itérations de l'algorithme itératif servant à obtenir les poids de redressement (les multiplicateurs de Lagrange et le critère d'arrêt qui est décrit dans la partie description).

Résultats

Statistiques descriptives (avant redressement) : dans ce tableau sont affichées pour toutes les modalités des variables auxiliaires les statistiques descriptives obtenues avant le redressement : l'effectif et le pourcentage dans l'échantillon, les sommes marginales dans la population (effectif et pourcentage).

Poids finaux : dans ce tableau sont affichés les poids redressés obtenus par la méthode sélectionnée. On peut aussi y visualiser les rapports de poids et les variables auxiliaires initiales.

Statistiques descriptives (après redressement) : dans ce tableau sont affichées pour toutes les modalités des variables auxiliaires les statistiques descriptives obtenues après le

redressement : l'effectif et le pourcentage dans l'échantillon, les sommes marginales dans la population (effectif et pourcentage).

Liste des combinaisons : dans ce tableau est affiché l'ensemble des combinaisons des modalités des variables auxiliaires présentes dans l'échantillon. Pour chaque combinaison, l'effectif et le rapport de poids sont donnés.

Détails des itérations : dans ce tableau sont affichés les détails des itérations de l'algorithme d'estimation des poids. Les lambdas sont les multiplicateurs de Lagrange et le critère d'arrêt est défini dans la partie description de cette aide.

Exemple

Un exemple de redressement d'un sondage est disponible sur le Centre d'aide XLSTAT à l'adresse suivante :

http://www.xlstat.com/demo-rakingf.htm

Bibliographie

Deming W.E. and Stephan F.F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *Annals of Mathematicals Statistics*, 11, 427-444.

Deville, J.-C., Särndal, C.-E. and Sautory, O. (1993). Generalized raking procedures in survey sampling. *Journal of the American Statistical Association*, vol. 88, no. 418, 376-382.

Créer un tableau de contingence

Utilisez cet outil pour créer un ou plusieurs tableaux de contingence à partir de deux ou plus variables qualitatives. Un test d'indépendance du Khi² peut être calculé.

Dans cette section :

Description

Un tableau de contingence est une manière efficace de résumer la relation entre deux variables qualitatives V_1 et V_2. Un tableau de contingence a la structure suivante :

$V_1 \setminus V_2$	**Modalité 1**	**…**	**Modalité j**	**…**	**Modalité m_2**
Modalité 1	$n(1,1)$		$n(1,j)$	…	$n(1,m_2)$
…	…	…	…	…	…
Modalité i	$n(i,1)$	…	$n(i,j)$	…	$n(i,m_2)$
…	…	…	…	…	…
Modalité m_1	$n(m_1,1)$	…	$n(m_1,j)$	…	$n(m_1,m_2)$

où $n(i,j) = n_{ij}$ est la fréquence des observations présentant à la fois la caractéristique i pour la variable V_1, et la caractéristique j pour la variable V_2.

Pour créer un tableau de contingence, la première transformation consiste en un recodage des deux variables qualitatives V_1 et V_2 en deux tableaux disjonctifs Z_1 et Z_2. Pour chaque modalité de la variable V_j, une colonne est crée dans Z_j. A chaque fois qu'une modalité m de la variable V_j correspond à un individu i, on affecte 1 à $Z_1(i,m)$. Les autres valeurs de Z_1 et Z_2 sont nulles. Le tableau de contingence des variables V_1 et V_2 n'est autre que le produit $Z_1' Z_2$ (où ' correspond à la transposition d'une matrice).

La distance du khi² a été proposée pour mesurer la distance entre les modalités. La somme de ces distances pour l'ensemble des cases du tableau donne la statistique du khi² qui suit asymptotiquement une loi du khi² à $(m_1 - 1)(m_2 - 1)$ degrés de liberté. Cette statistique permet de tester l'hypothèse d'indépendance entre les lignes et les colonnes du tableau de contingence.

La notion d'inertie inspirée de la physique est utilisée en Analyse Factorielle des Correspondances. L'inertie d'un nuage de points est la moyenne pondérée des carrés des

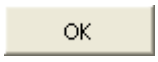
distances au centre de gravité. L'inertie totale du nuage des modalités est donnée par :

$$\phi^2 = \frac{\chi^2}{n} = \sum_{i=1}^{m_1}\sum_{j=1}^{m_2} \frac{\left(\frac{n_{ij}}{n} - \frac{n_{i.}n_{.j}}{n^2}\right)^2}{\frac{n_{i.}n_{.j}}{n^2}}, \text{ avec } n_{i.} = \sum_{j=1}^{m_2} n_{ij} \text{ et } n_{.j} = \sum_{j=1}^{m_1} n_{ij}$$

où n est la somme des fréquences du tableau de contingence. On voit ici que l'inertie totale est proportionnelle à la statistique du khi² de Pearson mesurée sur le tableau de contingence.

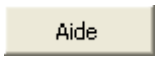
Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.



: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.

Onglet **Général**:

Variable(s) ligne : sélectionnez les données correspondant aux variables qualitatives qui seront les variables en ligne des tableaux de contingence créés. Si les libellés des variables ont été sélectionnés, veillez à ce que l'option « libellés des variables » soit bien activée.

Variable(s) colonne : sélectionnez les données correspondant aux variables qualitatives qui seront les variables en colonne des tableaux de contingence créés. Si les libellés des variables ont été sélectionnés, veillez à ce que l'option « libellés des variables » soit bien activée.

Analyse par groupe : activez cette option puis sélectionnez ici les données d'appartenance à des groupes si vous souhaitez que les calculs soient effectués sur chaque groupe séparément.

Poids : activez cette option si vous voulez pondérer les observations. Si vous n'activez pas cette option, les poids seront tous considérés comme valant 1. Les poids doivent être

impérativement supérieurs ou égaux à 0. Si un en- tête de colonne a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des variables : activez cette option si la première ligne des données sélectionnées contient un libellé.

Onglet **Options**:

Tri alphabétique des modalités : activez cette option pour que dans les divers résultats, les modalités soient triées alphabétiquement pour les deux variables sélectionnées.

Libellés Variable-Modalité : activez cette option pour que les libellés des lignes et des colonnes du tableau de contingence utilisent le nom de la variable suivi du nom des modalités. Si cette option n'est pas activée, les libellés sont construits uniquement à partir des noms des modalités.

Test du khi² : activez cette option pour effectuer le test du khi².

Niveau de signification (%) : entrez le niveau de signification à utiliser pour les différents tests (valeur par défaut : 5%).

Onglet **Données manquantes** :

Ne pas accepter les valeurs manquantes : activez cette option pour que XLSTAT empêche la poursuite des calculs si des valeurs manquantes sont détectées.

Supprimer les observations :

- **Suppression par paires** : activez cette option pour supprimer les observations comportant des données manquantes uniquement lorsque les variables impliquées dans les calculs comportent des données manquantes.
- **Pour tous les Y** : choisissez cette option pour supprimer toutes les observations pour lesquelles des données sont manquantes.

Regrouper les valeurs manquantes dans une nouvelle modalité : activez cette option pour regrouper les données manquantes dans une nouvelle modalité de la variable qualitative en question.

Onglet **Sorties**:

Liste des combinaisons : activez cette option pour afficher la liste des différentes combinaisons possibles des deux variables qualitatives, ainsi que les effectifs correspondants.

Tableau de contingence : activez cette option pour afficher le tableau de contingence.

Inertie par case : activez cette option pour afficher les inerties correspondant à chacune des cellules du tableau de contingence.

Khi² par case : activez cette option pour afficher les Khi² correspondant à chacune des cases du tableau de contingence.

Significativité par case : activez cette option pour afficher un tableau indiquant, pour chaque case, si la valeur observée est égale (=), inférieure (<) ou supérieure (>) à la valeur théorique, et pour effectuer un test (test exact de Fisher sur un tableau 2×2 ayant le même effectif total que le tableau complet, et les mêmes sommes marginales pour la case en question), afin de déterminer si l'écart à la valeur théorique est significatif ou non.

Les p-values associées sont aussi affichées.

Effectifs observés : activez cette option pour afficher le tableau des effectifs observés. Ce tableau est presque identique au tableau de contingence, la différence venant des sommes marginales pour les lignes et les colonnes.

Effectifs théoriques : activez cette option pour afficher le tableau des effectifs théoriques estimés à partir des sommes marginales.

Proportions ou pourcentages / Ligne : activez cette option pour afficher le tableau des proportions ou pourcentages par ligne qui correspondent aux effectifs observés divisés par les sommes marginales des lignes.

Proportions ou pourcentages / Colonne : activez cette option pour afficher le tableau des proportions ou pourcentages par colonne qui correspondent aux effectifs observés divisés par les sommes marginales des colonnes.

Proportions ou pourcentages / Total : activez cette option pour afficher le tableau des proportions ou pourcentages calculés comme les effectifs observés divisés par l'effectif total.

Synthèse pour tous les groupes : activez cette option pour afficher un résumé de l'ensemble des tableaux de contingence.

Onglet **Graphiques** :

Vue 3D du tableau de contingence / du tableau croisé : activez cette option pour afficher le diagramme en bâton en 3 dimensions correspondant au tableau de contingence ou au tableau croisé.

Tableau de contingence : activez cette option pour afficher le graphique associé au tableau de contingence.

Proportions ou pourcentages / Ligne : activez cette option pour afficher le graphique associé au tableau des proportions ou pourcentages par ligne.

Proportions ou pourcentages / Colonne : activez cette option pour afficher le graphique associé au tableau des proportions ou pourcentages par colonne.

Synthèse pour tous les groupes : activez cette option pour afficher les graphiques associés à chacun des groupes du tableau de synthèse.

Options des graphiques :

- **Type de graphique** :
  - **Groupé** : choisissez cette option pour afficher les graphiques sous forme de barres regroupées par modalité.
  - **Barres empilées** : choisissez cette option pour afficher les graphiques sous forme de barres empilées. Cela permet de comparer les effectifs ou les fréquences des sous-échantillons à ceux d'un échantillon complet.
- **Diagrammes en bâtons** :
  - **Effectifs** : choisissez cette option pour afficher l'effectif correspondant à chaque barre.
  - **Pourcentages** : choisissez cette option pour afficher le % de population correspondant à chaque barre

Exemple

Un exemple de création de tableaux de contingence est disponible sur leCentre d'aide XLSTAT à l'adresse suivante :

http://www.xlstat.com/demo-contf.htm

Tableaux disjonctifs complets

Utilisez cet outil pour créer un tableau disjonctif complet à partir d'une ou plusieurs variables qualitatives.

Dans cette section :

Description

Un tableau disjonctif consiste en l'éclatement d'un tableau défini par n observations et q variables qualitatives en un tableau défini par n observations et p indicatrices où p est la somme des nombres de modalités des q variables : chaque variable Q(j) est décomposée en un sous-tableau à q(j) colonnes où la colonne k contient des 1 pour les observations correspondant à la k-ième modalité et 0 pour les autres observations.

Boîte de dialogue

OK : cliquez sur ce bouton pour lancer les calculs.

Annuler : cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

Aide : cliquez sur ce bouton pour afficher l'aide.

 : cliquez sur ce bouton pour rétablir les options par défaut.

 : cliquez sur ce bouton pour effacer les sélections de données.

Données : sélectionnez les données sur la feuille Excel.

Libellés des variables : activez cette option si la première ligne des données sélectionnées (données et libellés des observations) contient un en- tête.

Libellés des observations : activez cette option si vous voulez utiliser des libellés d'observations disponibles sur une feuille Excel pour l'affichage des résultats. Si l'option « Libellés des variables » est activée, la première cellule de la sélection doit comprendre un en-tête. Si vous n'activez pas cette option, des libellés seront automatiquement créés (Obs1, Obs2, …).

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Statistiques descriptives : activez cette option pour calculer et afficher les statistiques descriptives des variables qualitatives sélectionnées.

Afficher l'en-tête du rapport : désactivez cette option si vous souhaitez que le tableau disjonctif complet commence dès la première ligne de la feuille Excel (cas d'une sortie dans une feuille ou un classeur), et non après l'en- tête du rapport.

Exemple

Tableau initial :

	Q1	**Q2**
Obs1	A	C
Obs2	B	D
Obs3	B	E
Obs4	A	D

Tableau disjonctif complet :

	Q1-1	**Q1-B**	**Q2-C**	**Q2-D**	**Q2-E**
Obs1	1	0	1	0	0
Obs2	0	1	0	1	0
Obs3	0	1	0	0	1
Obs4	1	0	0	1	0

Questions à réponses multiples

Utilisez cet outil pour transformer des tableaux qui incluent des questions à réponses multiples en un tableau où les questions à réponses multiples sont transformées de manière à pouvoir être analysées.

Dans cette section :

Description

Il est courant que les sondages comportent des questions à réponses multiples. Voici un exemple très simple : "Quelles sont vos couleurs préférées ?". Certaines personnes répondront "Bleu", d'autres "Bleu, Rouge", d'autres "Vert, Violet, Jaune". Toutes les combinaisons sont possibles à partir de la liste des possibilités qui sont données dans l'enquête. Le résultat de l'enquête est un tableau donnant les réponses de chaque individu à chaque question. Pour les questions à réponses multiples, ce sera généralement comme décrit ci-dessus, avec les différentes réponses séparées par un délimiteur.

Avec ce formatage, l'analyse des données et les techniques statistiques nécessitant une organisation des données structurée ne peuvent pas être appliquées. Une transformation des questions à réponses multiples est nécessaire. Les colonnes correspondant aux questions à réponses multiples sont remplacées par autant de colonnes qu'il y a de réponses dans les colonnes de départ (Note : si une réponse n'a jamais été cochée, elle n'aura pas de colonne dans le tableau restructuré'), avec au sein de chaque colonne, des réponses Oui/Non indiquant si les répondants ont coché ou non l'item.

Pour l'exemple ci-dessus nous avons:

|Couleurs| |--| |Bleu| |Bleu,Rouge| |Vert,Violet,Jaune|	qui est transformé en	|Bleu|Jaune|Rouge|Vert|Violet| |--|--|--|--|--| |Oui|Non|Non|Non|Non| |Oui|Non|Oui|Non|Non| |Non|Oui|Non|Oui|Oui|

Boîte de dialogue

OK : cliquez sur ce bouton pour lancer les calculs.

Annuler : cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

Aide : cliquez sur ce bouton pour afficher l'aide.

 : cliquez sur ce bouton pour rétablir les options par défaut.

 : cliquez sur ce bouton pour effacer les sélections de données.

 : cliquez sur ce bouton pour changer la façon dont vous chargez les données dans XLSTAT. Si le bouton possède une icône de souris, XLSTAT vous permet de faire une sélection à la souris des données. Si le bouton possède une icône de liste, XLSTAT vous permet de sélectionner les données à partir d'une liste. Si le bouton possède une icône orange, des boutons complémentaires avec un point d'interrogation  sont affichés pour vous permettre d'importer les données à partir de fichiers.

Données : sélectionnez toutes les données que vous souhaitez que XLSTAT analyse. Si des questions qui ne sont pas des questions à réponses multiples sont incluses, XLSTAT insérera les colonnes correspondant aux différentes réponses des questions à réponses multiples, entre les autres questions. Si des en-têtes de colonnes ont été sélectionnés, vérifiez que l'option "Libellés des colonnes" est bien activée.

Délimiteur : sélectionnez le délimiteur utilisé pour séparer les réponses multiples.

Confirmer les questions : activez cette option pour que XLSTAT vous demande de confirmer quelles questions sont des questions à réponses multiples dans les données.

Résultats

XLSTAT affiche le tableau des données restructurées. Si des questions qui ne sont pas des questions à réponses multiples ont été incluses, XLSTAT insère les colonnes correspondant aux différentes réponses des questions à réponses multiples, entre les autres questions, en respectant l'ordre dans le tableau initial.

Exemple

Un exemple d'utilisation de cet outil est disponible à l'adresse suivante :

http://www.xlstat.com/demo-mcrf.htm

Discrétisation

Utilisez cet outil pour discrétiser une variable numérique. Plusieurs choix de discrétisation sont proposés.

Dans cette section :

Description

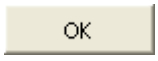
Discrétiser une variable numérique revient à la transformer en une variable ordinale. Ce procédé est très communément utilisé en marketing, où il est souvent appelé « segmentation ».

XLSTAT propose plusieurs méthodes de discrétisation plus ou moins automatiques. Le nombre de classes (ou intervalles, ou segments) générés est fixé soit par l'utilisateur (par exemple avec la méthode des amplitudes égales), soit par la méthode elle-même (par exemple, avec l'option 80-20, où deux classes sont créées).

L'algorithme de classification automatique de Fisher peut être très lent si le nombre de données dépasse le millier. Cette méthode génère un nombre de classes au plus égal au nombre de classes demandées, la méthode permettant de découvrir automatiquement que certaines classes peuvent être fusionnées.

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.

 : cliquez sur ce bouton pour effacer les sélections de données.

 : cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT vous permet de sélectionner les données par colonnes ou par plage. Si la flèche est vers la droite, XLSTAT vous permet de sélectionner les données par lignes ou par plage.

Onglet **Général**:

Tableau observations/variables : sélectionnez un tableau comprenant N objets décrits par P descripteurs. Si des en-têtes de colonnes ont été sélectionnés, veuillez vérifier que l'option « Libellés des variables » est activée. Si plusieurs variables sont sélectionnées, elles seront chacune à leur tour discrétisées.

Méthode : choisissez la méthode de discrétisation:

- **Amplitude constante** : choisissez cette méthode pour créer des classes de même amplitude. Entrez alors l'amplitude. Vous pouvez ensuite spécifier le minimum, correspondant à la borne inférieure de l'intervalle correspondant à la première classe. Cette valeur doit être inférieure ou égale au minimum de la série. Si le minimum n'est pas spécifié, la borne inférieure correspondra au minimum de la série.
- **Intervalles** : choisissez cette méthode pour créer un nombre donné d'intervalles de même amplitude. Entrez alors le nombre d'intervalles. L'amplitude des intervalles est déterminée à partir de la différence entre les maximum et minimum de la série Vous pouvez aussi spécifier le minimum, correspondant à la borne inférieure du premier intervalle. Cette valeur doit être inférieure ou égale au minimum de la série. Si le minimum n'est pas spécifié, la borne inférieure correspondra au minimum de la série.
- **Effectifs égaux** : choisissez cette méthode pour que les classes créées comprennent toutes le même nombre d'observations (dans la mesure du possible). Entrez alors le nombre d'intervalles (classes) à créer.
- **Automatique (Fisher)** : choisissez cette méthode pour créer les classes en utilisant l'algorithme de Fisher. Lorsque le nombre de données dépasse le millier, cet algorithme peut être très lent. Entrez alors le nombre d'intervalles (classes) à créer. Le nombre de classes créées peut être éventuellement inférieur à la valeur entrée, l'algorithme pouvant regrouper des classes non significativement différentes.
- **Automatique (k-means)** : choisissez cette méthode pour créer les intervalles en utilisant l'algorithme k-means. Entrez alors le nombre d'intervalles (classes) à créer.
- **Intervalles (définis par l'utilisateur)** : choisissez cette méthode pour sélectionner une colonne contenant en ordre croissant la borne inférieure du premier intervalle, et la borne supérieure de tous les intervalles.
- **80-20** : choisissez cette méthode pour créer deux classes, la première comprenant les 80 premiers % de la série, cette dernière étant classée en ordre croissant, la seconde

contenant les 20% restant.

- **20-80** : choisissez cette méthode pour créer deux classes, la première comprenant les 20 premiers % de la série, cette dernière étant classée en ordre croissant, la seconde contenant les 80% restant.
- **80-15-5 (ABC)** : choisissez cette méthode pour créer trois classes, la première comprenant les 80 premiers % de la série, cette dernière étant classée en ordre croissant, la seconde contenant les 15% suivant, et la troisième contenant les 5% restant. Cette classification est parfois appelées ABC.
- **5-15-80** : choisissez cette méthode pour créer trois classes, la première comprenant les 5 premiers % de la série, cette dernière étant classée en ordre croissant, la seconde contenant les 15% suivant, et la troisième contenant les 80% restant.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des variables : activez cette option si la première ligne des données sélectionnées (Tableau observations/variables, libellés des observations, poids) contient un libellé.

Libellés des observations : activez cette option si vous voulez utiliser des libellés d'observations pour l'affichage des résultats. Si l'option « Libellés des variables » est activée, la première cellule de la sélection doit comprendre un en-tête. Si vous n'activez pas cette option, des libellés seront automatiquement créés (Obs1, Obs2, …).

Afficher l'en-tête du rapport : désactivez cette option pour que l'en- tête du rapport ne soit pas affiché.

Onglet **Options**:

Poids : activez cette option si vous voulez pondérer les observations. Si vous n'activez pas cette option, les poids seront tous considérés comme valant 1. Les poids doivent être impérativement supérieurs ou égaux à 0. Si un en- tête de colonne a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

- **Standardiser les poids** : si vous activez cette option, les poids sont standardisés de telle sorte que leur somme soit égale au nombre d'observations.

Onglet **Données manquantes** :

Ne pas accepter les valeurs manquantes : activez cette option pour que XLSTAT empêche la poursuite des calculs si des valeurs manquantes sont détectées.

Supprimer les observations :

- **Pour l'échantillon correspondant** : activez cette option pour ne pas prendre en compte une observation dont l'une des données est manquante, uniquement pour les échantillons pour lesquels une donnée est manquante.
- **Pour tous les échantillons** : activez cette option pour ne pas prendre en compte une observation dont l'une des données est manquante, pour tous les échantillons sélectionnés.

Estimer les données manquantes : activez cette option pour estimer les données manquantes en utilisant la moyenne de l'échantillon.

Onglet **Sorties**:

Statistiques descriptives : activez cette option pour afficher les statistiques descriptives pour les échantillons sélectionnés.

Barycentres : activez cette option pour afficher les coordonnées des barycentres des classes.

Objets centraux : activez cette option pour afficher les coordonnées de l'objet le plus proche du barycentre de chaque classe.

Résultats par classe : activez cette option pour afficher un tableau donnant les statistiques et les objets correspondant à chacune des classes.

Résultats par objet : activez cette option pour afficher un tableau donnant pour chaque objet sa classe d'affectation dans l'ordre initial des objets.

Onglet **Graphiques** :

Histogrammes : activez cette option pour afficher les histogrammes des échantillons. Pour la distribution théorique, la fonction de densité est affichée.

- **Barres** : choisissez cette option pour afficher des histogrammes avec une barre pour chaque intervalle.
- **Lignes continues** : choisissez cette option pour afficher des histogrammes avec une ligne continue.

Fonction de répartition empirique : activez cette option pour afficher les histogrammes cumulés des échantillons. Pour la distribution théorique, la fonction de répartition est affichée.

- **Basés sur l'histogramme** : choisissez cette option pour afficher des histogrammes cumulés basés sur la même définition d'intervalles que les histogrammes.

- **Fonction de répartition empirique** : choisissez cette option pour afficher des histogrammes cumulés qui correspondent en réalité à la fonction de répartition empirique de l'échantillon.

Ordonnées des histogrammes : choisissez quelle grandeur doit être utilisée pour les histogrammes : densité, effectif ou fréquence.

Résultats

Statistiques simples : dans ce tableau sont affichés pour les variables sélectionnées, le nombre d'observations, le nombre de données manquantes, le nombre de données non manquantes, la moyenne, et l'écart-type.

Un **histogramme et la fonction de répartition empirique** sont affichés si les options correspondantes ont été activées. Les statistiques des différents intervalles sont affichées à la suite.

Barycentres des classes : dans ce tableau sont affichées les coordonnées des barycentres des classes pour les différents descripteurs.

Distances entre les barycentres des classes : dans ce tableau sont affichées les distances euclidiennes entre les barycentres des classes pour les différents descripteurs.

Objets centraux : dans ce tableau sont affichées pour chaque classe les coordonnées de l'objet le plus proche du barycentre de la classe.

Distances entre les objets centraux : dans ce tableau sont affichées les distances euclidiennes entre les objets centraux des classes pour les différents descripteurs.

Résultats par classe : les statistiques descriptives des classes (nombre d'objets, somme des poids, variance intra-classe, distance minimale au barycentre, distance maximale au barycentre, distance moyenne au barycentre) sont affichées dans la première partie du tableau. Dans la seconde partie sont affichés les objets.

Résultats par objet : dans ce tableau est indiquée, pour chaque objet, sa classe d'affectation dans l'ordre initial des objets.

Exemple

Un exemple de création de tableaux de contingence est disponible sur leCentre d'aide XLSTAT à l'adresse suivante :

http://www.xlstat.com/demo-discf.htm

Bibliographie

Arabie P., Hubert L.J. and De Soete G. (1996). Clustering and Classification. Wold Scientific, Singapore.

Everitt B.S., Landau S. and Leese M. (2001). Cluster Analysis (4th edition). Arnold, London.

Fisher W.D. (1958). On grouping for maximum homogeneity. *Journal of the American Statistical Association*, **53**, 789-798.

Gestion des données

Utilisez cet outil pour transformer des tableaux de données. Huit fonctions sont proposées : dédoublonner, grouper, joindre (interne et externe), filtrer (garder et supprimer) et empiler/désempiler. Ces méthodes sont communes dans les systèmes de gestion de base de données, mais ne sont pas proposées par Excel.

Dans cette section :

Description

Dédoublonner

Il est parfois nécessaire de dédoublonner un tableau de données : certaines observations peuvent être présentes plusieurs fois (on parle alors de doublons) suite à la fusion de plusieurs sources de données, ou suite à des erreurs de saisie.

Grouper

Le groupement est utile lorsque vous voulez agréger des données. Imaginez par exemple le cas d'un tableau contenant des enregistrements de ventes (une colonne pour l'identifiant client, et une colonne avec le montant de la vente) que vous voudriez agréger pour avoir une ligne par client, avec l'identifiant du client et le montant total des ventes pour ce client. XLSTAT vous permet d'obtenir ce tableau en quelques secondes. La somme n'est que l'une des six possibilités proposées.

Joindre

La jointure est une opération courante en gestion de base de données. Elle permet de fusionner « horizontalement » deux tables sur la base d'une information commune dénommée la clef. Par exemple, imaginez que vous avez mesuré quelques indicateurs chimiques sur 150 sites. Ensuite, vous voulez ajouter l'information géographique sur ces mêmes sites où les données ont été recueillies. Votre table d'informations géographiques contient l'information sur 1000 sites, y compris les 150 sites étudiés. Afin d'éviter le travail fastidieux de fusionner manuellement les deux tables, une jointure permet d'obtenir en quelques secondes la table fusionnée qui comprend à la fois les données recueillies et l'information géographique. On distingue deux types de jointure :

- Jointure interne : la table fusionnée comprend uniquement les clefs communes aux deux tables de départ.

- Jointure externe : la table fusionnée comprend une ligne par clef, qu'elle soit présente dans une seule des tables de départ ou dans les deux.

Filtrer (Garder/Supprimer)

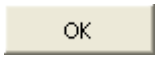
Cet outil vous permet de sélectionner un tableau et de créer un nouveau tableau qui comprend (Garder) ou exclut (Supprimer) les lignes pour lesquelles la valeur, dans une colonne donnée, correspond à une valeur contenue dans une liste sélectionnée par l'utilisateur.

Empiler/Désempiler

Cet outil permet de transformer un tableau organisé sous forme d'une colonne par groupe en deux colonnes l'une associée à la valeur de la variable et la seconde au groupe associé. L'opération inverse (désempiler) est aussi possible. Ceci permet par exemple de transformer des données sous forme de colonnes en données adaptées à une analyse de la variance à un facteur.

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT vous permet de sélectionner les données par colonnes ou par plage. Si la flèche est vers la droite, XLSTAT vous permet de sélectionner les données par lignes ou par plage.


: cliquez sur ce bouton pour changer la façon dont vous chargez les données dans XLSTAT. Si le bouton possède une icône de souris, XLSTAT vous permet de faire une sélection à la souris des données. Si le bouton possède une icône de liste, XLSTAT vous permet de sélectionner les données à partir d'une liste. Si le bouton possède une icône orange, des

boutons complémentaires avec un point d'interrogation  sont affichés pour vous permettre d'importer les données à partir de fichiers.

Onglet **Général**:

Données : ce champ n'est visible que si les méthodes « Dédoublonner », « Grouper », « Filtrer » ou « Empiler » sont activées. Sélectionnez le tableau des données que vous voulez dédoublonner, grouper, filtrer ou empiler. Si des en-têtes de colonnes ont été sélectionnés, veuillez vérifier que l'option « Libellés des variables » est activée.

Libellés des observations : ce champ n'est visible que si la méthode « Dédoublonner » est activée. Activez cette option pour sélectionner les libellés d'observations qui seront ensuite utilisés pour l'affichage des résultats. Si l'option « Libellés des variables » est activée, la première cellule de la sélection doit comprendre un en-tête. Si vous n'activez pas cette option, des libellés seront automatiquement créés (Obs1, Obs2, …).

Deviner les types : cette option n'est visible que si la méthode « Grouper » est activée. Activez cette option si vous souhaitez que XLSTAT devine le type des variables sélectionnées (numérique ou nominal). Si cette option n'est pas activée, XLSTAT vous demandera de confirmer ou de modifier les types des variables.

Tableau 1 : ce champ n'est visible que si la méthode « Jointure » est activée. Sélectionnez les données correspondant à la première table de jointure. Si des en-têtes de colonnes ont été sélectionnés, veuillez vérifier que l'option « Libellés des variables » est activée.

Tableau 2 : ce champ n'est visible que si la méthode « Jointure » est activée. Sélectionnez les données correspondant à la première table de jointure. Si des en-têtes de colonnes ont été sélectionnés, veuillez vérifier que l'option « Libellés des variables » est activée.

Méthode : choisissez la méthode de gestion de données à utiliser :

- Dédoublonner
- Grouper
- Jointure (Interne)
- Jointure (Externe)
- Filtrer (Garder/Supprimer)
- Empiler
- Désempiler

Opération : cette option n'est visible que si la méthode « Grouper » est activée. Choisissez l'opération à appliquer lors de l'agrégation des données. Pour les variables nominales, le mode est utilisé comme résultat.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des variables : activez cette option si la première ligne des données sélectionnées (Tableau observations/variables, libellés des observations, poids) contient un libellé.

Afficher l'en-tête du rapport : désactivez cette option si vous souhaitez que le tableau des résultats commence dès la première ligne de la feuille Excel (cas d'une sortie dans une feuille ou un classeur), et non après l'en- tête du rapport.

Onglet **Données manquantes**:

Cet onglet n'est visible que pour les méthodes « Dédoublonner » ou « Grouper ».

Ne pas accepter les valeurs manquantes : activez cette option pour que XLSTAT empêche la poursuite des calculs si des valeurs manquantes sont détectées.

Supprimer les observations : activez cette option pour supprimer les observations comportant des données manquantes.

Ignorer les données manquantes : si vous activez cette option, XLSTAT ignorera les données manquantes.

Onglet **Sorties**:

Cet onglet n'est visible que pour les méthodes « Dédoublonner » ou « Grouper ».

Statistiques descriptives : activez cette option pour calculer et afficher les statistiques descriptives des variables sélectionnées.

Les options suivantes ne sont visibles que pour la méthode « Dédoublonner » :

Tableau dédoublonné : activez cette option pour afficher le tableau dédoublonné.

Effectifs : activez cette option pour afficher dans la dernière colonne du tableau la fréquence de chaque observation dans le tableau initial (1 correspond à une donnée non dupliquée).

Doublons : activez cette option pour afficher les données présentes au moins deux fois dans le tableau initial.

Résultats

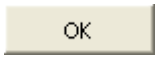
Les résultats sont affichés à l'endroit souhaité. En fonction de la méthode choisie les résultats peuvent comprendre une ou plusieurs sortie.

Nettoyer les données textuelles

Utilisez cet outil pour supprimer les espaces à gauche et/ou à droite de chaînes de caractères ou corriger des répétitions d'espace ou remplacer des chaînes de caractères.

Dans cette section :

Boîte de dialogue



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.


: cliquez sur ce bouton pour changer la façon dont vous chargez les données dans XLSTAT. Si le bouton possède une icône de souris, XLSTAT vous permet de faire une sélection à la souris des données. Si le bouton possède une icône de liste, XLSTAT vous permet de sélectionner les données à partir d'une liste. Si le bouton possède une icône orange, des boutons complémentaires avec un point d'interrogation  sont affichés pour vous permettre d'importer les données à partir de fichiers.

Onglet **Général** :

Source de données : choisissez la source de votre texte parmi ces deux options : * **Feuille de calcul** : sélectionnez les données sur la feuille Excel (une ligne par document). * **Fichiers** : sélectionnez un ou plusieurs fichiers textes (version WINDOWS) ou le dossier qui les contient (version MAC).

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des colonnes : activez cette option si la première ligne des données sélectionnées comprend un libellé qui ne doit pas être traité.

Afficher l'en-tête du rapport : désactivez cette option si vous souhaitez que le tableau des résultats commence dès la première ligne de la feuille Excel (cas d'une sortie dans une feuille ou un classeur), et non après l'en- tête du rapport.

Onglet **Options** :

Supprimer les espaces : activez cette option pour supprimer les espaces à gauche et/ou à droite des données textuelles.

- **A gauche** : activez cette option pour supprimer les espaces à gauche.
- **A droite** : activez cette option pour supprimer les espaces à droite.
- **Y compris ** : activez cette option pour supprimer également les caractères ' '; typiques de textes provenant de pages internet.

Espaces entre les mots : activez cette option puis choisissez le nombre d'espaces maximum souhaité entre deux mots.

Remplacer les caractères : activez cette option pour remplacer des caractères (ou des mots) par d'autres caractères (ou mots) * **Remplacer** : sélectionnez une colonne avec les caractères (ou mots) à trouver et à remplacer. * **par** : sélectionnez une colonne avec les caractères (ou mots) de remplacement.

NB : il s'agit d'un tableau de codage entré en deux étapes, les deux colonnes doivent alors avoir le même nombre de lignes. Les remplacements sont fait dans l'ordre d'entrée et la casse est prise en compte. Si des en-têtes ont été sélectionnés, veuillez vérifier que l'option « Libellés des colonnes » est activée.

Résultats

Les résultats sont affichés à l'endroit souhaité. Le tableau contient les chaînes traitées par les méthodes choisies. Le tableau de codage est affiché si l'option "Remplacer les caractères" est activée.

Codage

Utilisez cet outil pour recoder un tableau en utilisant un tableau de codage comprenant les valeurs initiales et les codes qui doivent les remplacer dans le nouveau tableau.

Dans cette section :

Boîte de dialogue

OK : cliquez sur ce bouton pour lancer les calculs.

Annuler : cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

Aide : cliquez sur ce bouton pour afficher l'aide.

 : cliquez sur ce bouton pour rétablir les options par défaut.

 : cliquez sur ce bouton pour effacer les sélections de données.

Données : sélectionnez les données sur la feuille Excel. Si des en-têtes ont été sélectionnés, veuillez vérifier que l'option « Libellés des colonnes » est activée.

Tableau de codage : sélectionnez deux colonnes correspondant au tableau de codage. La première colonne doit contenir les valeurs telles qu'elles sont dans le tableau des données sélectionnées, et la seconde colonne les codes correspondants à utiliser dans le tableau recodé. Si des en-têtes ont été sélectionnés, veuillez vérifier que l'option « Libellés des colonnes » est activée.

Libellés des colonnes : activez cette option si la première ligne des données sélectionnées (données et tableau de codage) contient un libellé.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Afficher l'en-tête du rapport : désactivez cette option si vous souhaitez que le tableau disjonctif complet commence dès la première ligne de la feuille Excel (cas d'une sortie dans une feuille ou un classeur), et non après l'en- tête du rapport.

Exemple

Un exemple de création de tableaux de contingence est disponible sur le Centre d'aide XLSTAT à l'adresse suivante :

http://www.xlstat.com/demo-codef.htm

Codage présence/absence

Utilisez cet outil pour transformer un tableau de listes (ou attributs) en un tableau de présence/absence indiquant les fréquences des différents éléments pour chacune des listes.

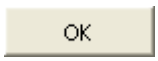
Dans cette section :

Description

Cet outil permet par exemple de transformer un tableau contenant p colonnes correspondant à p listes d'objets en un tableau à p lignes et q colonnes, où q est le nombre d'objets différents contenu dans les p listes, et où pour chaque cellule du tableau, on a 1 si l'objet est présent et 0 s'il est absent.

Par exemple, dans le domaine de l'écologie, si on a p relevés d'espèces avec en colonne, pour chaque relevé, les différentes espèces trouvées, on obtiendra un tableau croisé indiquant la présence ou l'absence de chacune des espèces pour chacun des relevés.

Boîte de dialogue



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.

Données : sélectionnez les données sur la feuille Excel. Si des en-têtes ont été sélectionnés, veuillez vérifier que l'option « Libellés des colonnes » est activée.

Libellés des colonnes : activez cette option si la première ligne des données sélectionnées contient un libellé.

Codage présence/absence par :

- **Lignes** : choisissez cette option si chaque ligne correspond à une liste.

- **Colonnes** : choisissez cette option si chaque colonne correspond à une liste.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Afficher l'en-tête du rapport : désactivez cette option si vous souhaitez que le tableau de présence/absence commence dès la première ligne de la feuille Excel (cas d'une sortie dans une feuille ou un classeur), et non après l'en-tête du rapport.

Exemple

Tableau initial :

Liste 1	**Liste 2**
E1 E1 E2 E1 E3	E3 E1 E4

Tableau de présence/absence :

	E1	**E2**	**E3**	**E4**
Liste 1	1	1	1	0
Liste 2	1	0	1	1

Codage en rangs

Utilisez cet outil pour recoder un tableau à n observations et p variables quantitatives en un tableau contenant le rang des valeurs, les rangs étant déterminés variable par variable.

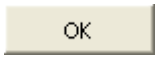
Dans cette section :

Description

Cet outil vous permet de recoder un tableau à n observations et p variables quantitatives en un tableau contenant le rang des valeurs, les rangs étant déterminés variable par variable. Le codage en rang peut vous permettre de convertir un tableau de variables quantitatives continues en un tableau de variables quantitatives discrètes, si seule la relation d'ordre est intéressante et non les valeurs elles-mêmes.

Deux stratégies sont possibles pour la prise en compte des ex aequo : soit on leur affecte un rang moyen, soit on leur affecte le rang le plus faible.

Boîte de dialogue



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.

Données : sélectionnez les données sur la feuille Excel. Si des en-têtes ont été sélectionnés, veuillez vérifier que l'option « Libellés des colonnes » est activée.

Libellés des colonnes : activez cette option si la première ligne des données sélectionnées (données et libellés des observations) contient un libellé.

Libellés des observations : activez cette option si vous voulez utiliser des libellés d'observations disponibles sur une feuille Excel pour l'affichage des résultats. Si l'option « Libellés des colonnes » est activée, la première cellule de la sélection doit comprendre un en-tête. Si vous n'activez pas cette option, des libellés seront automatiquement créés (Obs1, Obs2, …).

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Tenir compte des ex aequo : activez cette option pour tenir compte de la présence d'ex aequo et pour adapter en conséquence le rang des valeurs ex aequo.

- **Rangs moyens** : choisissez cette option pour remplacer le rang des valeurs ex aequo par la moyenne des rangs.
- **Minimum** : choisissez cette option pour remplacer le rang des valeurs ex aequo par le minimum de leur rang.

Afficher l'en-tête du rapport : désactivez cette option si vous souhaitez que le tableau échantillonné commence dès la première ligne de la feuille Excel (cas d'une sortie dans une feuille ou un classeur), et non après l'en- tête du rapport.

Exemple

Tableau initial :

	V1	**V2**
Obs1	1.2	12
Obs2	1.6	11
Obs3	1.2	10
Obs4	1.4	10.5

Tableau recodé en rangs (rang moyen pour les ex aequo) :

	R1	R2
Obs1	1.5	4
Obs2	4	3
Obs3	1.5	1
Obs4	3	2

Tableau recodé en rangs (rang le plus faible pour les ex aequo) :

	R1	R2
Obs1	1	4
Obs2	4	3
Obs3	1	1
Obs4	3	2

Importer un fichier de données

Utilisez cet outil pour charger les données d'un fichier texte en mémoire. Cela permet ainsi de charger des données au-delà des limites d'une feuille de calcul.

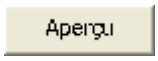
Dans cette section :

Description

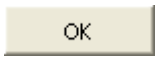
Il est parfois nécessaire de traiter des données stockées sous forme de fichier texte. Celles-ci peuvent représenter des quantités de lignes et de colonnes supérieures aux limites imposées par Excel dans une feuille de calcul. Pour cela, XLSTAT permet de charger ces données dans la mémoire (seule une fonctionnalité le permet pour le moment, voir Gestion des données).

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la lecture des données que pour le décryptage du fichier. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour afficher un aperçu des dix premières colonnes du fichier.



: cliquez sur ce bouton pour enregistrer les paramètres de lecture du fichier.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans enregistrer.


: cliquez sur ce bouton pour rétablir les options par défaut.

Onglet **Général** :

Chemin d'accès : ce champ permet de définir le fichier à charger dans XLSTAT.

Format : choisissez le type de fichier à charger :

- CSV Files : Tous types de fichiers plat écrit avec l'extension *.csv
- Text Files : Tous types de fichiers plat écrit avec l'extension *.txt
- All Files : Tous types de fichiers plat écrit avec toutes extensions

Séparateur : ce champ permet de définir le caractère qui sert de séparateur de colonnes dans le fichier.

Séparateur décimal : ce champ définit le type de séparateur décimal utilisé pour les valeurs numériques.

Codage de caratère : choisissez le type d'encodage à utiliser :

- UTF-8 : ce codage considére que les caractères du fichier sont codés sur 8 bits. Cette fonctionnalité vérifie s'il s'agit d'un encodage UTF-8 avec BOM (Byte Order Mark) et retire cet ensemble caractère si c'est le cas.
- UTF-16 : ce codage suppose que les caractères du fichier sont codés sur 16 bits. Cette fonctionnalité lit le BOM (Byte Order Mark) et définit s'il s'agit d'un fichier écrit en big-endian (BE) ou little-endian (LE). Par défaut, en l'absence du BOM, le codage est de type little-endian (LE).

Marqueur de texte : ce champ permet de définir un caractère englobant. Ainsi, le fichier de données peut contenir des éléments complexes tel que des éléments composés de deux mots séparés d'un espace. Par exemple, le séparateur choisi est ESPACE et un des titres de colonnes est "Liste de mots". Si le marqueur de texte ' " ' est bien défini, le terme "Liste de mots" entouré de guillemets sera vu comme un seul élément.

Libellés des variables : activez cette option pour sélectionner les libellés des variables qui seront ensuite utilisés pour l'affichage des résultats. Si elle est activé, les éléments de la première ligne sont considérés comme les libellés de chaque colonne. Si vous n'activez pas cette option, des libellés seront automatiquement créés (Var1, Var2, …).

Libellés des observations : activez cette option pour sélectionner les libellés d'observations qui seront ensuite utilisés pour l'affichage des résultats. Si l'option « Libellés des observations » est activée, le premier élément de chaque ligne sera désigné comme le premier. Si vous n'activez pas cette option, des libellés seront automatiquement créés (Obs1, Obs2, …).

Onglet **Options** :

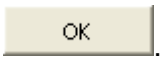
Lire à partir de la ligne : ce champ permet de définir la ligne à partir de laquelle on lit le fichier et ainsi de ne pas inclure des potentielles premières lignes dans les données.

Symbole de commentaire : ce champ définit le caractère de commentaire. Tous les éléments d'une ligne situés à la droite de ce caractère sont ignorés. Ainsi, le jeu de données peut se composer d'éléments de commentaire pour une meilleure lisibilité du fichier sans que cela ait des conséquences pour la lecture des données. Par exemple, sur Mac, certains utilisateurs ont pour habitude de mettre un en-tête commencant par le caractère ' # ' dans leur fichier de données pour préciser la date de génération du fichier, l'auteur, ...

Détection du format : activez cette option pour que les paramètres soient définis automatiquement lors de la sélection du fichier dans **Chemin d'accès**.

Bouton **Aperçu** :

Celui-ci fait apparaître une fenêtre d'aperçu dans laquelle on retrouve les champs de l'onglet **Général**. En plus de ceux-ci, la fenêtre contient également le champ **Ligne(s) affichée(s)** qui permet de définir le nombre de lignes que l'on souhaite pré-visualiser. De cette manière, il est possible d'avoir une pré-visualisation dynamique du chargement des données en même temps

que l'on change les valeurs des différents champs. Pour valider ces paramètres, il suffit de cliquer sur le bouton 

.

Description des données

Statistiques descriptives et Graphiques univariés

Utilisez cet outil pour calculer des statistiques descriptives et afficher des graphiques univariés (Box plots, Scattergrams, …) pour un ensemble de variables quantitatives et/ou qualitatives.

Dans cette section :

Description

Avant d'utiliser des méthodes d'analyse avancées comme par exemple une analyse discriminante ou une régression multiple, il est nécessaire dans un premier temps, de découvrir les données afin d'identifier des tendances, de repérer des anomalies ou tout simplement de disposer d'informations essentielles telles que le minimum, le maximum, ou la moyenne d'un échantillon de données.

XLSTAT vous propose un nombre important de statistiques descriptives et de graphiques qui vous permettront d'avoir un premier aperçu pertinent de vos données.

Bien que vous puissiez sélectionner plusieurs variables (ou échantillons) à la fois, XLSTAT calcule l'ensemble des statistiques descriptives pour chacun des échantillons indépendamment.

Statistiques descriptives pour les données quantitatives :

Soit un échantillon composé de N données quantitatives $\{y_1, y_2, \ldots y_N\}$, dont les poids respectifs sont $\{W_1, W_2, \ldots W_N\}$.

- **Nombre d'observations** : le nombre N de données dans l'échantillon sélectionné.
- **Nombre de données manquantes** : le nombre de données manquantes dans l'échantillon analysé. Pour le calcul des statistiques qui suivent, les données identifiées comme manquantes sont ignorées. On définit par **n** le nombre de données non manquantes, et par $\{x_1, x_2, \ldots x_n\}$ le sous-échantillon des données non manquantes dont les poids respectifs sont $\{w_1, w_2, \ldots w_n\}$.

- **Somme des poids** *: la somme des poids, notée W. Lorsque tous les poids valent 1, ou lorsque les poids sont « standardisés », on a $W = n$.
- **Répartition par sous-échantillon (%)** *: la répartition de chaque sous-échantillon.
- **Minimum** : le minimum de la série analysée.
- **Maximum** : le maximum de la série analysée.
- **Fréquence du minimum** * : la fréquence du minimum de la série.
- **Fréquence du maximum** * : la fréquence du maximum de la série.
- **Amplitude** : l'amplitude est la différence entre le maximum et le minimum de la série.
- **1-er quartile** * : le premier quartile **Q1** est défini comme la valeur telle que 25% des données lui sont inférieurs.
- **Médiane** * : la médiane **Q2** est telle que 50% des données lui sont inférieurs.
- **3-ème quartile** * : le troisième quartile **Q3** est défini comme la valeur telle que 75% des données lui sont inférieurs.
- **Somme** * : la somme pondérée des données est définie par :

$$S = \sum_{i=1}^{n} w_i x_i$$

- **Moyenne** * : la moyenne de l'échantillon est définie par :

$$\hat{\mu} = S/W$$

- **Variance n** * : la variance de l'échantillon est définie par :

$$s_n^2 = \frac{\sum_{i=1}^{n} w_i (x_i - \hat{\mu})^2}{W}$$

Remarque 1 : lorsque tous les poids valent 1, la variance est la somme des écarts quadratiques à la moyenne, divisée par n, d'où la dénomination.

Remarque 2 : la variance n est une estimation biaisée de la variance, qui suppose que l'échantillon est bien représentatif de la population totale. La variance n-1 est calculée au contraire en tenant compte d'une approximation liée à l'échantillonnage.

- **Variance n-1** * : la variance estimée de l'échantillon définie par :

$$s_{n-1}^2 = \frac{\sum_{i=1}^{n} w_i (x_i - \hat{\mu})^2}{W - W/n}$$

Remarque 1 : lorsque tous les poids valent 1, la variance est la somme des écarts quadratiques à la moyenne, divisée par n-1, d'où la dénomination.

Remarque 2 : la variance n est une estimation biaisée de la variance, qui suppose que l'échantillon est bien représentatif de la population totale. La variance n-1 est calculée au contraire en tenant compte d'une approximation liée à l'échantillonnage.

- **Ecart-type n** * : l'écart-type de l'échantillon défini par $s_n = \sqrt{s_n^2}$.
- **Ecart-type n-1** * : l'écart-type estimé de l'échantillon défini par $s_{n-1} = \sqrt{s_{n-1}^2}$.
- **Coefficient de variation** * : ce coefficient n'est calculé que si la moyenne de l'échantillon n'est pas nulle. Il est défini par :

$$CV = \frac{s_n}{\hat{\mu}}$$

Ce coefficient mesure la dispersion d'un échantillon relativement à sa moyenne. Il permet de comparer la dispersion d'échantillons dont les échelles ou les moyennes sont sensiblement différentes.

- **Asymétrie (Pearson)** * : le coefficient d'asymétrie de Pearson est défini par :

$$g_1 = \frac{\hat{\mu_3}}{s_{n-1}^3} \text{ avec } \hat{\mu_3} = \frac{\sum_{i=1}^{n} w_i (x_i - \hat{\mu})^3}{W}$$

Ce coefficient, appelé *skewness* en anglais, donne une indication quant à la forme de la distribution de l'échantillon. Dans le cas d'une valeur négative (respectivement positive) la distribution est concentrée à gauche (respectivement à droite) de la moyenne.

- **Asymétrie (Fisher)** * : le coefficient d'asymétrie de Fisher est défini par :

$$G_1 = g_1 \frac{\sqrt{W(W - W/n)}}{W - 2W/n}$$

Contrairement au précédent, ce coefficient est non biaisé sous hypothèse de normalité des données. Ce coefficient donne une indication quant à la forme de la distribution de l'échantillon. Dans le cas d'une valeur négative (respectivement positive) la distribution est concentrée à gauche (respectivement à droite) de la moyenne.

- **Asymétrie (Bowley)** * : le coefficient d'asymétrie de Bowley est défini par :

$$A(B) = \frac{Q_1 - 2Q_2 + Q_3}{Q_3 - Q_1}$$

- **Aplatissement (Pearson ou excess)** * : le coefficient d'aplatissement de Pearson est défini par :

$$g_2 = \frac{\hat{\mu_4}}{s_{n-1}^4} - 3 \text{ avec } \hat{\mu_4} = \frac{\sum_{i=1}^{n} w_i (x_i - \hat{\mu})^4}{W}$$

Ce coefficient appelé en anglais *kurtosis* ou parfois *excess kurtosis* donne une indication quant à la forme de la distribution de l'échantillon. Dans le cas d'une valeur négative (respectivement

positive), le pic de la distribution de l'échantillon est plus (respectivement moins) aplati que celui d'une loi normale.

- **Aplatissement (Population)** * : le coefficient d'aplatissement de Fisher est défini par :

$$G_2 = \frac{(W - W/n)}{(W - 2W/n)(W - 3W/n)}((W + W/n)g2 + 6)$$

Contrairement au précédent, ce coefficient est non biaisé sous hypothèse de normalité des données. Ce coefficient appelé en anglais *kurtosis* ou parfois *excess kurtosis* donne une indication quant à la forme de la distribution de l'échantillon. Dans le cas d'une valeur négative (respectivement positive), le pic de la distribution de l'échantillon est plus (respectivement moins) aplati que celui d'une loi normale.

- **Ecart-type de la moyenne** * : cette statistique est définie par :

$$s_\mu = \frac{s_n}{\sqrt{n-1}}$$

- **Borne inf. de la moyenne (x% ou niveau de signification α=1-x/100)** * : cette statistique correspond à la borne inférieure de l'intervalle de confiance à x% autour de la moyenne. Cette statistique est définie par :

$$L_\mu = \hat{\mu} - s_\mu |t_{(a/2)}|$$

- **Borne sup. de la moyenne (x% ou niveau de signification α=1-x/100)** * : cette statistique correspond à la borne supérieure de l'intervalle de confiance à x% autour de la moyenne. Cette statistique est définie par :

$$U_\mu = \hat{\mu} + s_\mu |t_{(\alpha/2)}|$$

- **Ecart-type de la variance** * : cette statistique est définie par :

$$s_{s^2} = s^2_{n-1}\sqrt{\frac{2}{W-1}}$$

- **Borne inf. de la variance (x% ou niveau de signification α=1-x/100)** * : cette statistique correspond à la borne inférieure de l'intervalle de confiance à x% autour de la variance. Cette statistique est définie par :

$$L_{s^2} = s_\sigma / \chi_{(1-\alpha/2)}$$

- **Borne sup. de la variance (x% ou niveau de signification α=1-x/100)** * : cette statistique correspond à la borne supérieure de l'intervalle de confiance à x% autour de la variance. Cette statistique est définie par :

$$U_{s^2} = s_\sigma / \chi_{(\alpha/2)}$$

- **Ecart-type (Asymétrie (Fisher))** *: l'écart-type du coefficient d'asymétrie de Fisher est défini par :

$$se(G_1) = \sqrt{\frac{6W(W-1)}{(W-2)(W+1)(W+3)}}$$

- **Ecart-type (Aplatissement (Population))** *: l'écart-type du coefficient d'aplatissement de Fisher est défini par :

$$se(G_2) = 2se(G1)\sqrt{\frac{(W^2-1)}{(W-3)(W+5)}}$$

- **Ecart absolu moyen** *: comme l'écart-type ou la variance, ce coefficient mesure la dispersion (ou variabilité) de l'échantillon. Il est défini par :

$$e(\mu) = \frac{\sum_{i=1}^{n} w_i|x_i - \mu|}{W}$$

- **Ecart absolu médian** *: cette statistique correspond à la médiane des écarts absolus à la médiane.

- **Moyenne géométrique** *: cette statistique n'est calculée que si toutes les données sont strictement positives. Elle est définie par :

$$\mu_G = exp(\frac{1}{W}\sum_{i=1}^{n} w_i ln(x_i))$$

Si tous les poids sont égaux à 1, on a

$$\mu_G = \sqrt[n]{\prod_{i=1}^{n} x_i}$$

- **Ecart-type géométrique** * : cette statistique est définie par :

$$\sigma_G = exp(\frac{1}{W}\sum_{i=1}^{n} w_i [ln(x_i) - ln(\mu_G)]^2)$$

- **Moyenne harmonique** * : cette statistique est définie par :

$$\mu_H = \frac{W}{\sum_{i=1}^{n} \frac{w_i}{x_i}}$$

(*) Les statistiques suivies d'un astérisque tiennent compte du poids des observations.

Statistiques descriptives pour les données qualitatives :

Pour un échantillon composé de N données qualitatives, on définit :

- **Nombre d'observations** : le nombre N de données dans l'échantillon sélectionné.

- **Nombre de données manquantes** : le nombre de données manquantes dans l'échantillon analysé. Pour le calcul des statistiques qui suivent, les données identifiées

comme manquantes sont ignorées. On définit par n le nombre de données non manquantes, et par {w1, w2, … wn} les poids des données non manquantes.

- **Somme des poids** *: la somme des poids, notée W. Lorsque tous les poids valent 1, on a $W = n$.
- **Mode** *: le mode de l'échantillon analysé. Autrement dit, la modalité la plus fréquente.
- **Fréquence du mode** *: la fréquence de la modalité à laquelle correspond le mode
- **Modalité** : le nom des différentes modalités présentes dans l'échantillon.
- **Fréquence par modalité** *: la fréquence de chacune des modalités.
- **Fréquence relative par modalité** *: la fréquence relative de chacune des modalités.
- **Borne inf. des fréquences (x% ou niveau de signification α=1-x/100)** * : cette statistique correspond à la borne inférieure de l'intervalle de confiance à x% autour de chaque fréquence.
- **Borne sup. des fréquences (x% ou niveau de signification α=1-x/100)** * : cette statistique correspond à la borne supérieure de l'intervalle de confiance à x% autour de chaque fréquence.
- **Proportion par modalité** * : la proportion de chacune des modalités.
- **Borne inf. des proportions (x% ou niveau de signification α=1-x/100)** * : cette statistique correspond à la borne inférieure de l'intervalle de confiance à x% autour de chaque proportion.
- **Borne sup. des proportions (x% ou niveau de signification α=1-x/100)** * : cette statistique correspond à la borne supérieure de l'intervalle de confiance à x% autour de chaque proportion.

(*) Les statistiques suivies d'un astérisque tiennent compte du poids des observations.

Plusieurs types de graphiques sont disponibles pour les données quantitatives et les données qualitatives :

Graphiques pour les données quantitatives :

- **Box plots** : ces représentations univariées d'échantillons de données quantitatives sont parfois appelées « diagrammes boîtes à moustaches ». C'est une représentation simple et assez complète puisque dans la version proposée par XLSTAT sont affichés le minimum, le 1er quartile, la médiane, la moyenne, le 3ème quartile, ainsi que les deux limites (les extrémités des « moustaches ») au-delà desquelles on peut considérer que les valeurs sont anormales. La moyenne est affichée sous la forme d'un + rouge, et la médiane sous la forme d'une ligne noire. Les limites sont ainsi calculées :

Limite inférieure : $L_{inf} = X(i)$ tel que $\{X(i) - [Q1 - 1.5(Q3 - Q1)]\}$ soit minimal et $X(i) = Q1 - 1.5(Q3 - Q1)$.

Limite supérieure : $L_{sup} = X(i)$ tel que $\{X(i) - [Q3 + 1.5(Q3 - Q1)]\}$ soit minimal et $X(i) = Q3 + 1.5(Q3 - Q1)$.

Les valeurs en dehors de l'intervalle $]Q1 - 3(Q3 - Q1); Q3 + 3(Q3 - Q1)[$ sont affichées avec un symbole *; et les valeurs comprises dans $[Q1 - 3(Q3 - Q1); Q1 - 1.5(Q3 - Q1)]$ ou $[Q3 + 1.5(Q3 - Q1); Q3 + 3(Q3 - Q1)]$ sont affichées avec un symbole "o".

XLSTAT permet de produire également des box plots « entaillés » (*notched* box plots en anglais). Les extrémités de l'entaille permettent de visualiser un intervalle de 95% autour de la médiane. Les extrémités sont calculées suivant la formule suivante :

- **Limite inférieure** : $N_{inf} = Médiane - [1.58(Q3 - Q1)]/\sqrt{(}n)$
- **Limite supérieure** : $N_{sup} = Médiane + [1.58(Q3 - Q1)]/\sqrt{(}n)$

Ces formules données par McGill *et al.* (1978) dérivent de la normalité asymptotique de la distribution de médiane dans le cadre de comparaisons de médiane. Si les tailles d'échantillon sont homogènes, les box plots entaillés permettent de comparer les médianes et la variabilité de l'échantillon (du fait de la largeur de l'intervalle).

XLSTAT permet également de faire varier la largeur des box plots en fonction de la de la racine carrée taille de l'échantillon représenté.

- **Scattergrams** : ces représentations univariées permettent de donner une idée de la distribution et de la pluralité éventuelle des modes d'un échantillon. Tous les points sont représentés, ainsi que la moyenne et la médiane.
- **Strip plots** : ces diagrammes représentent sous forme de bandes (*strip* en anglais) les données de l'échantillon. Sur un intervalle donné, plus les bandes sont serrées ou épaisses plus il y a de données.
- **Stem-and-leaf plots** : ces représentations univariées, aussi appelées diagrammes branches et feuilles, permettent de visualiser la distribution des données tout en conservant une vision précise des valeurs, contrairement aux histogrammes. Chaque donnée est scindée en deux parties : une partie branche, à gauche du diagramme et une partie feuille, à droite du diagramme. Pour reconstituer une donnée, il suffit de multiplier le nombre [branche virgule feuille] par l'unité affichée au-dessus de la représentation.
- **Graphiques P-P (loi normale)** : les graphiques Probabilité-Probabilité (*P-P plots* en anglais) permettent de comparer la fonction de répartition empirique d'un échantillon à celle d'un échantillon distribué suivant une loi normale de même moyenne et même variance. Si l'échantillon suit une loi normale, les points doivent être confondus avec la première bissectrice du plan.
- **Graphiques Q-Q (loi normale)** : les graphiques Quantile-Quantile (*Q-Q plots* en anglais) permettent de comparer les quantiles de l'échantillon à ceux d'un échantillon distribué suivant une loi normale de même moyenne et même variance. Si l'échantillon suit une loi normale, les points doivent être confondus avec la première bissectrice du plan.

- **Graphique des moyennes** : les graphiques des moyennes représentent, sous forme de diagrammes en bâtons, les moyennes de chacune des variables. Il est aussi possible d'afficher sur ces graphiques les **barres d'erreurs** sous trois formes différentes :
  - l'écart-type défini par $s_n = \sqrt{s_n^2}$.
  - l'erreur standard définie par $err = \frac{s_n}{\sqrt{n}}$.
  - l'intervalle de confiance défini par $L_\mu = \hat{\mu} \pm s_\mu |t_{(a/2)}|$.

Graphiques pour les données qualitatives :

Diagrammes en bâtons : activez cette option pour représenter sous forme de diagrammes en bâtons les effectifs ou les fréquences des différentes modalités des variables qualitatives.

Diagrammes en secteurs : activez cette option pour représenter sous forme de diagrammes en secteurs (ou camemberts) les effectifs ou les fréquences des différentes modalités des variables qualitatives.

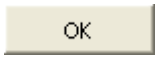
Diagrammes en secteurs doubles : ces graphiques permettent de comparer les effectifs ou les fréquences de sous-échantillons à ceux d'un échantillon complet.

Anneaux : cette option n'est active que si une colonne de sous- échantillons a été sélectionnée. Ces graphiques permettent de comparer les effectifs ou les fréquences des sous-échantillons à ceux d'un échantillon complet.

Barres empilées : cette option n'est active que si une colonne de sous- échantillons a été sélectionnée. Ces graphiques permettent de comparer les effectifs ou les fréquences des sous-échantillons à ceux d'un échantillon complet.

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les

variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.

Onglet **Général** :

Données quantitatives : activez cette option pour sélectionner les échantillons de données quantitatives pour lesquels vous voulez calculer les statistiques descriptives. Si des en-têtes ont été sélectionnés, veuillez vérifier que l'option « Libellés des échantillons » est activée.

Données qualitatives : activez cette option pour sélectionner les échantillons de données qualitatives pour lesquels vous voulez calculer les statistiques descriptives. Si des en-têtes ont été sélectionnés, veuillez vérifier que l'option « Libellés des échantillons » est activée.

Sous-échantillons : activez cette option pour sélectionner une colonne indiquant les noms ou les indices des sous-échantillons correspondant à chacune des observations.

- Libellés Variable-Modalité : activez cette option pour utiliser des libellés longs pour l'affichage des résultats. Les libellés Variable-Modalité sont composés du nom de la variable comme préfixe, et de la modalité du sous-échantillon comme suffixe.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des échantillons : activez cette option si la première ligne des sélections (données quantitatives, qualitatives, sous-échantillons, poids) contient un libellé.

Poids : activez cette option si vous voulez pondérer les observations. Si vous n'activez pas cette option, les poids seront tous considérés comme valant 1. Les poids doivent être impérativement supérieurs ou égaux à 0. Si un en- tête de colonne a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

- **Standardiser les poids** : si vous activez cette option les poids sont standardisés de telle sorte que leur somme soit égale au nombre d'observations.

Onglet **Options** :

Statistiques descriptives : activez cette option pour calculer et afficher les statistiques descriptives.

Graphiques : activez cette option pour afficher des graphiques.

Normaliser : activez cette option pour centrer-réduire les données avant de procéder à l'analyse.

Remettre à l'échelle de 0 à 100 : activez cette option remettre les données à l'échelle en faisant en sorte que le minimum soit 0 et le maximum 100.

Comparer à l'échantillon total : cette option n'est active que si une colonne de sous-échantillons a été sélectionnée. Activez cette option pour que les statistiques descriptives et les graphiques soient aussi affichés pour l'échantillon total.

Onglet **Sorties** :

Données quantitatives : activez les options pour les statistiques descriptives que vous voulez calculer. Les différentes statistiques sont présentées dans la section description.

- **Toutes** : cliquez sur ce bouton pour tout sélectionner.
- **Aucune** : cliquez sur ce bouton pour tout désélectionner.
- **Affichage vertical** : activez cette option pour que le tableau des statistiques descriptives soit affiché verticalement (une ligne par statistique descriptive).

Données qualitatives : activez les options pour les statistiques descriptives que vous voulez calculer. Les différentes statistiques sont présentées dans la section description.

- **Toutes** : cliquez sur ce bouton pour tout sélectionner.
- **Aucune** : cliquez sur ce bouton pour tout désélectionner.
- **Affichage vertical** : activez cette option pour que le tableau des statistiques descriptives soit affiché verticalement (une ligne par statistique descriptive).

Onglet **Graphiques (1)** :

Cet onglet concerne les données quantitatives.

Sous-onglet **Types de graphiques** :

Box plots : activez cette option pour afficher les box plots (ou graphiques boîtes et moustaches). Voir la section description pour plus de détails.

Scattergrams : activez cette option pour afficher les scattergrams. La moyenne (+ rouge) et la médiane (trait rouge) sont systématiquement affichées.

Strip plots : activez cette option pour afficher les strip plots. Sur ces graphiques, une bande correspond à une observation.

Stem-and-leaf plots : activez cette option pour afficher les stem-and- leaf plots (ou diagrammes branches et feuilles).

- **Unité** : 10^ : activez cette option si vous souhaitez configurer manuellement l'unité de subdivision des données en branches et en feuilles.

Graphiques P-P (loi-normale) : activez cette option pour afficher les graphiques P-P.

Graphiques Q-Q (loi-normale) : activez cette option pour afficher les graphiques Q-Q.

Graphiques des moyennes : activez cette option pour afficher les graphiques des moyennes.

Sous-onglet **Options** :

Cet onglet concerne les options pour les box plots, les scattergrams et les strip plots.

Horizontaux : activez cette option pour afficher des box plots, scattergrams et strip plots horizontaux.

Verticaux : activez cette option pour afficher des box plots, scattergrams et strip plots verticaux.

Grouper les graphiques : activez cette option pour regrouper sur un même graphique les différents box plots, scattergrams et strip plots de manière à pouvoir les comparer.

Il est possible de spécifier la « **Dimension** » correspondant au nombre maximum de box plots à regrouper. Par défaut, ce nombre est **automatiquement** choisi en fonction du nombre de variables et de modalités. Vous pouvez aussi le spécifier manuellement. Ce nombre est au maximum de 20 avec Excel 2003 et de 40 pour Excel 2007 et au-delà.

- **Modalités** : activez cette option pour grouper les modalités des sous-échantillons si vous en avez sélectionné.
- **Variables** : activez cette option pour grouper les variables y compris dans le cas où il y a des sous-échantillons. Vous pouvez alors afficher une **ligne grise** entre les variables.
- **Tri par moyenne** : activez cette option pour trier les variables et modalités par moyenne, de la plus élevée à la plus petite.

Entaillés : activez cette option pour afficher des box plots entaillés (*notched box plots*).

Adapter la largeur : activez cette option pour que la largeur des box plots dépende de la taille de l'échantillon.

Minimum/Maximum : activez cette option pour systématiquement afficher les points correspondant au minimum et au maximum (box plots).

Valeurs extrêmes : activez cette option pour afficher les points correspondant aux valeurs extrêmes (box plots) avec un cercle évidé.

Position des étiquettes : choisissez la position des étiquettes sur les graphiques verticaux. Elles peuvent être soit en bas, soit en haut, soit alternativement en bas et en haut.

Légende : activez cette option pour afficher les symboles des statistiques utilisées dans les box plots.

Colorer l'intérieur : activez cette option pour colorer l'intérieur des box plots.

Colorer par groupe : activez cette option pour colorer les box plots en fonction des sous-échantillons définis dans l'onglet Général.

Onglet **Graphiques (2)** :

Cet onglet concerne les données qualitatives.

Diagrammes en bâtons : activez cette option pour représenter sous forme de diagrammes en bâtons les effectifs ou les fréquences des différentes modalités des variables qualitatives.

Diagrammes en secteurs : activez cette option pour représenter sous forme de diagrammes en secteurs (ou camemberts) les effectifs ou les fréquences des différentes modalités des variables qualitatives.

- **Doubles** : cette option n'est active que si une colonne de sous-échantillons a été sélectionnée. Ces graphiques permettent de comparer les effectifs ou les fréquences des sous-échantillons à ceux d'un échantillon complet.

Anneaux : cette option n'est active que si une colonne de sous- échantillons a été sélectionnée. Ces graphiques permettent de comparer les effectifs ou les fréquences des sous-échantillons à ceux d'un échantillon complet.

Barres empilées : cette option n'est active que si une colonne de sous- échantillons a été sélectionnée. Ces graphiques permettent de comparer les effectifs ou les fréquences des sous-échantillons à ceux d'un échantillon complet.

Valeurs utilisées : choisissez le type de données à afficher :

- **Effectifs** : choisissez cette option pour que l'échelle des graphiques corresponde aux effectifs des modalités.
- **Fréquences** : choisissez cette option pour que l'échelle des graphiques corresponde aux fréquences des modalités.

Exemple

Un exemple de calcul de statistiques descriptives et de génération de box plots est disponible sur le Centre d'aide XLSTAT à l'adresse suivante :

http://www.xlstat.com/demo-bpf.htm

Bibliographie

Filliben J.J. (1975). The probability plot correlation coefficient Test for normality. *Technometrics*, **17** (1), 111-117.

DeCarlo L.T. (1997). On the meaning and Use of Kurtosis. *Psychological Methods*, **2** (3), 292-307.

McGill R., Tukey J. W. and Larsen W. A. (1978). Variations of box plots. *The American Statistician*, **32**, 12-16.

Sokal R.R. and Rohlf F.J. (1995). Biometry. The Principles and Practice of Statistics in Biological Research. Third Edition. Freeman, New York.

Tomassone R., Dervin C. and Masson J.P. (1993). Biométrie. Modélisation de Phénomènes Biologiques. Masson, Paris.

Velleman P. F. and Hoaglin D. C. (1981). Applications, Basics and Computing of Exploratory Data Analysis, Duxbury, Belmont, CA.

Caractérisation de variables

Utilisez cet outil pour caractériser des éléments (variables quantitatives, qualitatives ou modalités de variables qualitatives) nommés « éléments à caractériser », en explorant les liaisons qu'ils entretiennent avec des éléments caractérisants (variables quantitatives, qualitatives ou modalités de variables qualitatives). Pour cela différents tests statistiques (paramétriques ou non paramétriques) sont utilisés.

Dans cette section :

Description

Les différentes caractérisations de variables proposées dans cet outil sont :

1. Caractérisation d'une variable quantitative :

1-1. par des variables quantitatives :

La caractérisation d'une variable quantitative par d'autres variables quantitatives s'effectue par l'intermédiaire du coefficient de corrélation. Pour chaque variable quantitative caractérisante, on teste si celui-ci est significativement différent de 0 via le test de corrélation de Pearson (paramétrique) ou Spearman (non paramétrique). Plus le coefficient de corrélation est significativement différent de 0, plus les 2 variables (quantitatives) sont liées.

1-2. par des variables qualitatives :

La caractérisation d'une variable quantitative par des variables qualitatives s'effectue à l'aide de tests statistiques paramétriques ou non paramétriques. Si la p-value du test est inférieure à un seuil choisi on rejettera l'hypothèse d'indépendance entre les deux variables. Dans le cas paramétrique le test utilisé est le test de Fisher (comme en ANOVA). Dans le cas non paramétrique si la variable qualitative possède $k = 2$ modalités, on utilise le test de Mann-Whitney, si la variable qualitative possède plus de 2 modalités, on utilise le test de Kruskal-Wallis.

1-3. par des modalités d'une variable qualitative :

La caractérisation d'une variable quantitative par des modalités s'effectue à l'aide d'un indicateur appelé valeur test (Lebart, 2000). La valeur test $t_k(X)$ d'une variable quantitative X associée à la modalité k d'une variable qualitative est définie comme suit :

$$t_k(X) = \frac{\overline{X_k} - \overline{X}}{s_k(X)}$$

avec :

$$s_k^2(X) = \frac{n - n_k}{n - 1} \frac{s^2(X)}{n_k}$$

où $s^2(X)$ est la variance empirique de la variable X.
On calcule ensuite une p-value associée à cette valeur test, plus la p-value est proche de 0, plus la moyenne de la variable X sur la modalité k est différente de la moyenne générale.

1. Caractérisation d'une variable qualitative (à k modalités) :

2-1. par des variables quantitatives :

La caractérisation d'une variable qualitative par des variables quantitatives s'effectue à l'aide de tests statistiques paramétriques ou non paramétriques. Si la p-value du test est inférieure à un seuil choisi on rejettera l'hypothèse d'indépendance entre les deux variables. Dans le cas paramétrique le test utilisé est le test de Fisher (comme en ANOVA). Dans le cas non paramétrique si la variable qualitative possède $k = 2$ modalités, on utilise le test de Mann-Whitney, si la variable qualitative possède plus de 2 modalités, on utilise le test de Kruskal-Wallis.

2-2. par d'autres variables qualitatives :

La caractérisation d'une variable qualitative par d'autres variables qualitatives s'effectue à l'aide d'un test d'indépendance. Pour chaque variable qualitative caractérisante, on teste donc l'indépendance avec la variable qualitative à caractériser via le test d'indépendance du Chi² (paramétrique) ou le test exact de Fisher (non paramétrique).

1. Caractérisation d'une modalité d'une variable qualitative :

3-1. par des variables quantitatives :

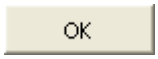
La caractérisation d'une modalité par des variables quantitatives s'effectue à l'aide de la valeur test comme expliqué en 1-3.

3-2. par d'autres modalités :

La caractérisation d'une modalité par d'autres modalités s'effectue à l'aide de la valeur test pour les variables qualitatives (Lebart, 2000) et de sa p-value associée. Une modalité est considérée comme caractéristique de la classe si son abondance dans la classe est jugée significativement supérieure à ce qu'on peut attendre compte tenu de sa présence dans la population. en notant n_{jk} le nombre d'individu ayant la modalité j parmi les n_k individus de la classe k, n_j le nombre d'individus ayant la modalité j et n l'effectif total, l'abondance de la modalité j est définie en comparant son pourcentage dans la $k$ème classe $\frac{n_{kj}}{n_k}$ à son pourcentage dans la population $\frac{n_j}{n}$.

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.

Onglet **Général**:

Y/ Eléments à caractériser :

Variable(s) Quantitative(s) : activez cette option si vous voulez caractériser une ou plusieurs variables quantitatives. Sélectionnez alors la ou les variables correspondantes sur la feuille Excel. Les données sélectionnées doivent être de type numérique. Si des en-têtes de colonnes ont été sélectionnés, veuillez vérifier que l'option "Libellés des variables" est activée.

Variable(s) Qualitative(s) : activez cette option si vous voulez caractériser une ou plusieurs variables qualitatives. Sélectionnez alors la ou les variables correspondantes sur la feuille Excel. Activez aussi cette option si vous voulez caractériser des modalités de variables qualitatives. Les données sélectionnées peuvent être de tout type, mais les données numériques sont automatiquement considérées comme nominales. Si des en-têtes de colonnes ont été sélectionnés, veuillez vérifier que l'option "Libellés des variables" est activée.

Modalités : activez cette option si vous voulez caractériser les modalités des variables qualitatives sélectionnées précédemment.

X/ Eléments caractérisants :

Variable(s) Quantitative(s) : activez cette option si vous voulez qu'une ou plusieurs variables quantitatives soient caractérisantes. Sélectionnez alors la ou les variables correspondantes sur la feuille Excel. Les données sélectionnées doivent être de type numérique. Si des en-têtes de colonnes ont été sélectionnés, veuillez vérifier que l'option "Libellés des variables" est activée.

Variable(s) Qualitative(s) : activez cette option si vous voulez qu'une ou plusieurs variables qualitatives soient caractérisantes. Sélectionnez alors la ou les variables correspondantes sur la feuille Excel. Activez aussi cette option si vous voulez que des modalités de variables qualitatives soient caractérisantes. Les données sélectionnées peuvent être de tout type, mais les données numériques sont automatiquement considérées comme nominales. Si des en-têtes de colonnes ont été sélectionnés, veuillez vérifier que l'option "Libellés des variables" est activée.

Modalités : activez cette option si vous voulez que les modalités des variables qualitatives sélectionnées précédemment soient caractérisantes.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des variables : activez cette option si la première ligne des sélections (données quantitatives, qualitatives, sous-échantillons, poids) contient un libellé.

Poids des observations : activez cette option si vous voulez pondérer les observations. Si vous n'activez pas cette option, les poids seront tous considérés comme valant 1. Pour les tests non paramétriques les poids doivent être des entiers positifs. Si un en-tête de colonne a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

Onglet **Options**:

Filtrer les éléments caractérisants : Activez cette option si vous voulez filter les éléments caractérisants à afficher. Plusieurs options de filtrage sont disponibles, en fonction de l'option choisie, vous devez choisir un seuil pour les p-values (ou valeurs test) à afficher ou bien un nombre p d'éléments caractérisants à afficher.

Trier les éléments caractérisants : Activez cette option si vous voulez trier l'affichage des éléments caractérisants en fonction des p-values.

Niveau de signification : Entrez dans la cellule associée le niveau de signification que vous voulez.

Tests paramétrique : Activez cette option si vous souhaitez qu'un test paramétrique soit mis en œuvre.

Tests non naramétriques : Activez cette option si vous souhaitez qu'un test non paramétrique soit mis en œuvre.

Onglet **Données manquantes** :

Supprimer les observations : activez cette option pour ne pas prendre en compte une observation dont l'une des données est manquante.

Estimer les données manquantes : activez cette option pour estimer les données manquantes en utilisant la moyenne de l'échantillon ou la plus proche observation.

Onglet **Sorties**:

Statistiques descriptives : activez cette option pour afficher les statistiques descriptives des échantillons.

Graphique des p-values : activez cette option pour afficher le graphique des p-values.

Graphique des valeurs test : activez cette option pour afficher le graphique des valeurs test.

Résultats

Statistiques descriptives : le tableau de statistiques descriptives présente pour toutes les variables sélectionnées des statistiques simples.

En fonction des éléments à caractériser et des éléments caractérisants, le tableau de résultats affiché sera différent. Cependant quel que soit le cas, la p-value sera toujours affichée.

Si vous avez sélectionné l'option **Graphique des p-values** un graphique de type diagramme en bâtons représentant la valeur des p-values est également affiché en dessous de chaque tableau.

Exemple

Un exemple de caractérisation de variables est disponible sur le Centre d'aide XLSTAT à l'adresse suivante :

http://www.xlstat.com/demo-demodf.htm

Bibliographie

Lebart L., Morineau A. and Piron M. (2000) . Statistique Exploratoire Multidimensionnelle. Dunod, 181-184.

Morineau A. (1984) . Note sur la Caractérisation Statistique d'une Classe et les Valeurs-tests. Bulletin Technique du Centre de Statistique et d'Informatique Appliquées, 2(1-2), 20-27.

Estimation des quantiles

Utilisez cet outil pour calculer des quantiles associés à des variables quantitatives et afficher les propriétés qui leurs sont associées.

Dans cette section :

Description

Les quantiles sont des quantités qui peuvent être très utiles en statistique. Un quantile est obtenu à partir de la distribution de probabilité cumulée associée à une variable quantitative. On appelle généralement centile (ou percentile) le quantile ramené sur une échelle de 0 à 100.

XLSTAT vous propose cinq méthodes de calcul des quantiles, ainsi que deux techniques pour calculer des intervalles de confiance à partir des quantiles.

Bien que vous puissiez sélectionner plusieurs variables (ou échantillons) à la fois, XLSTAT calcule les quantiles pour chacun des échantillons indépendamment.

Définition des quantiles :

Soit une variable aléatoire X, nous utilisons comme notation la suivante : le p-quantile à 95% sera noté 0,95-quantile.

Soit $0 < p < 1$, le p-quantile associé à la variable aléatoire X est :

$$P(X \leq x) \geq p \text{ et } P(X \geq x) \geq 1 - p$$

Les quantiles sont des mesures utiles parce qu'elles sont moins sensibles aux distributions allongées et aux valeurs aberrantes.

Méthodes de calcul des quantiles :

XLSTAT permet d'utiliser cinq méthodes afin de calculer des quantiles.

Soit un échantillon composé de N données quantitatives $\{x_1, x_2, \dots x_N\}$, dont les poids respectifs sont $\{w_1, w_2, \dots w_N\}$. Nous noterons $x(1), \dots, x(N)$ et $w(1), \dots, w(N)$ après avoir ordonné les données.

Le p-quantile est noté y, on note j partie entière de $N \times p$ et g représente la part fractionnelle, on a : $g = N \times p \ - \ j$.

Nous avons donc :

Méthode de la moyenne pondérée à $x(Np)$:

$$y = (1 - g)x_{(j)} + gx_{(j+1)}$$

où x(0) est remplacé par x(1).

Méthode de l'observation la plus proche de $x(N * p)$:

$$\begin{aligned} &y = x_{(j)} \text{ if } g < 1/2 \\ &y = x_{(j)} \text{ if } g = 1/2 \text{ et j est pair} \\ &y = x_{(j+1)} \text{ if } g = 1/2 \text{ et j est impair} \\ &y = x_{(j+1)} \text{ if } g > 1/2 \end{aligned}$$

Fonction de distribution empirique :

$$\begin{aligned} &y = x_{(j)} \text{ si } g = 0 \\ &y = x_{(j+1)} \text{ si } g > 0 \end{aligned}$$

Méthode de la moyenne pondérée à $x((N + 1)p)$: dans ce cas, on prend $(N + 1)p = j + g$

$$y = (1 - g)x_{(j)} + gx_{(j+1)}$$

où x(N+1) est remplacé par x(N).

Fonction de distribution empirique avec mise à la moyenne :

$$\begin{aligned} &y = \tfrac{1}{2}(x_{(j)} + x_{(j+1)}) \text{ si } g = 0 \\ &y = x_{(j+1)} \text{ si } g > 0 \end{aligned}$$

Lorsque des poids sont associés aux données, une seule méthode est disponible :

$$y = \begin{cases} x_{(1)} & \text{si } w_{(1)} > pW \\ \frac{1}{2}(x_{(i)} + x_{(i+1)}) & \text{si } \sum_{j=1}^{i} w_{(j)} = pW \\ x_{(i+1)} \text{ si } \sum_{j=1}^{i} w_{(j)} < pW < \sum_{j=1}^{i+1} w_{(j)} \end{cases}$$

où $w(i)$ est le poids associé à $x(i)$ et $W = \sum_{j=1}^{N} w_j$.

Méthodes de calcul des intervalles de confiance :

Il est possible d'obtenir des intervalles de confiance pour les quantiles, deux types d'intervalles sont disponibles dans XLSTAT :

Intervalle de confiance en supposant la normalité des données :

L'intervalle de confiance avec une confiance de $100(1 - alpha)\%$ pour le p-quantile est :

$$[Np + Z_{\alpha/2}\sqrt{Np(1-p)} + 0.5;\ Np + Z_{1-\alpha/2}\sqrt{Np(1-p)} + 0.5]$$

Ce type d'intervalle de confiance peut être utilisé lorsque le nombre d'observations est grand (plus de 20) et si on suppose que les données suivent une distribution normale.

L'intervalle de confiance sans distribution :

L'intervalle de confiance avec une confiance de $100(1 - alpha)\%$ pour le p-quantile est :

$$[x_{(l)}; x_{(u)}]$$

l et u sont des entiers dont les valeurs sont symétriques autour de $[np] + 1$ avec $[np]$ partie entière de $n \times p$. On choisit $x(l)$ et $x(u)$ de façon à ce qu'ils soient le plus proche possible de $x([n + 1]p)$ tout en satisfaisant la formule suivante :

$$Q(u - 1, n, p) - Q(l - 1, n, p) \geq 1 - \alpha$$

où $Q(k, n, p)$ est probabilité binomiale cumulée :

$$Q(k, n, p) = \sum_{i=1}^{k} \left(\begin{array}{c} n \\ i \end{array} \right) p^i (1 - p)^{n-i}$$

Il est à noter que les intervalles de confiance ne peuvent pas être calculés lorsque des poids sont présents.

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.

OK : cliquez sur ce bouton pour lancer les calculs.

Annuler : cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

Aide : cliquez sur ce bouton pour afficher l'aide.

 : cliquez sur ce bouton pour rétablir les options par défaut.

 : cliquez sur ce bouton pour effacer les sélections de données.

 : cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.

Onglet **Général**:

Données : activez cette option pour sélectionner les échantillons de données quantitatives pour lesquels vous voulez calculer les statistiques descriptives. Si des en-têtes ont été sélectionnés, veuillez vérifier que l'option « Libellés des échantillons » est activée.

Méthode de calcul des quantiles : sélectionnez la méthode de calcul des quantiles que vous désirez utiliser (voir la partie description de l'aide d'XLSTAT pour plus de détails). La méthode par défaut est l'utilisation de la moyenne pondérée.

Intervalle de confiance :

- **Basé sur la distribution normale :** activez cette option si vous désirez afficher des intervalles de confiance sur les quantiles basés sur la distribution normale. Voir la section description pour plus de détails.
- **Sans hypothèse de distribution :** activez cette option si vous désirez afficher des intervalles de confiance sur les quantiles sans hypothèse de distribution. Voir la section description pour plus de détails.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des échantillons : activez cette option si la première ligne des sélections (données quantitatives, qualitatives, sous-échantillons, poids) contient un libellé.

Poids : activez cette option si vous voulez pondérer les observations. Si vous n'activez pas cette option, les poids seront tous considérés comme valant 1. Les poids doivent être impérativement supérieurs ou égaux à 0. Si un en- tête de colonne a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

Sous-échantillons : activez cette option pour sélectionner une colonne indiquant les noms ou les indices des sous-échantillons correspondant à chacune des observations.

Onglet **Données manquantes** :

Supprimer les observations : activez cette option pour ne pas prendre en compte une observation dont l'une des données est manquante.

Estimer les données manquantes : activez cette option pour estimer les données manquantes en utilisant la moyenne de l'échantillon.

Onglet **Sorties**:

Statistiques descriptives : activez cette option pour afficher les statistiques descriptives des échantillons.

Onglet **Graphiques** :

Fonction de répartition empirique : activez cette option pour afficher les histogrammes cumulés des échantillons. Pour la distribution théorique, la fonction de répartition est affichée.

Box plots : activez cette option pour afficher les box plots (ou graphiques boîtes et moustaches). Sur ces graphiques sont notamment affichés la médiane (trait rouge), le premier $(Q1)$ et le troisième $(Q3)$ quartiles (extrémités de la boîte) et les limites à partir desquelles on peut considérer qu'il s'agit de données potentiellement anormales. La limite inférieure est égale à $Q1 - 1,5 \times (Q3 - Q1)$, et la limite supérieure est égale à $Q3 + 1,5 \times (Q3 - Q1)$.

Scattergrams : activez cette option pour afficher les scattergrams. La médiane (trait rouge) est systématiquement affichée.

Afficher le quantile dans les graphiques (%) : cette option permet de rentrer un percentile et d'obtenir la valeur qui lui est associé ainsi qu'une représentation dans les différents graphiques sélectionnés.

Résultats

Statistiques simples : dans ce tableau sont affichées pour tous les échantillons les statistiques descriptives suivantes : le nombre d'observations, le nombre de données manquantes, le nombre de données non manquantes, la moyenne, et l'écart-type (non biaisé).

Tableau des quantiles : dans ce tableau sont affichés les percentiles pour les valeurs communément utilisées (0, 1, 5, 10, 25, 50, 75, 90, 95, 99 et 100). Pour chaque valeur, les intervalles de confiance associés sont aussi affichés.

Exemple

Un exemple de calcul des quantiles est disponible sur le Centre d'aide XLSTAT à l'adresse suivante :

http://www.xlstat.com/demo-quaf.htm

Bibliographie

Evans M., Hastings N. and Peacock B. (2000). Statistical Distribution. 3^{rd} edition, Wiley, New York.

Hahn J.H. and Meeker W.Q. (1991). Statistical intervals: A guide for Practitioners. Wiley, New York.

Histogrammes

Utiliser cet outil pour créer un histogramme à partir d'un échantillon de données quantitatives continues ou discrètes.

Dans cette section :

Description

L'histogramme est l'un des outils de visualisation les plus utilisés car il permet d'avoir très rapidement une idée de la distribution d'un échantillon de données quantitatives continues ou discrètes.

Définition des intervalles

L'un des enjeux pour la création d'un histogramme est la définition des intervalles, car pour un jeu de données déterminé, l'allure de l'histogramme en dépend entièrement. Entre les deux extrêmes de l'intervalle unique comprenant toutes les données et donnant une seule barre, et de l'histogramme où il y a un intervalle par donnée, il existe autant d'histogrammes possibles que de partitions des données.

Afin d'obtenir un résultat visuellement et/ou opérationnellement satisfaisant, la définition des intervalles peut nécessiter plusieurs aller-retours.

La méthode la plus classique consiste à utiliser des intervalles de même amplitude, la valeur du premier intervalle étant déterminée par la valeur minimale ou une valeur légèrement inférieure.

Afin de faciliter l'obtention d'histogrammes, XLSTAT vous propose de créer vos histogrammes soit en définissant le nombre d'intervalles, soit en définissant leur amplitude, soit en spécifiant vous-même les intervalles. Les intervalles sont considérés comme étant fermés pour la borne inférieure et ouverts pour la borne supérieure.

Histogramme cumulé

XLSTAT vous permet de créer des histogrammes cumulés qui correspondent soit au cumul des valeurs de l'histogramme, soit à la fonction de répartition empirique. L'utilisation de la fonction

de répartition empirique est recommandée pour une comparaison à une fonction de répartition d'une distribution théorique.

Comparaison à une distribution théorique

XLSTAT vous permet de comparer, si vous le souhaitez, l'histogramme à une distribution théorique dont vous pouvez fixer les paramètres. Néanmoins, si vous souhaitez vérifier si un échantillon est distribué suivant une loi donnée, vous pouvez utiliser l'outil d'ajustement d'une loi de distribution pour estimer les paramètres de la loi et éventuellement vérifier si l'hypothèse est acceptable.

XLSTAT permet l'utilisation des lois suivantes :

- Arcsinus (a) : la densité de cette loi (dérivée de la loi Bêta de type I) est donnée par :

$$f(x) = \frac{\sin(\pi\alpha)}{\pi x}(\frac{x}{1-x})^{\alpha-1}, \quad \text{avec } 0 < \alpha < 1, x \in [0,1]$$

On a $E(X) = \alpha$ et $V(X) = \alpha(1-\alpha)/2$

- Bernoulli (p) : la densité de cette loi est donnée par :

$$P(X=1) = p, P(X=0) = 1-p, \quad \text{avec } p \in [0,1]$$

On a $E(X) = p$ et $V(X) = p(1-p)$

La loi de Bernoulli, du nom du mathématicien suisse Jacob Bernoulli (1654-1705), permet de décrire les phénomènes aléatoires binaires où seuls deux événements peuvent survenir avec des probabilités respectives de p et $1-p$.

- Bêta (a, b) : la densité de cette loi (aussi appelée Bêta de type I) est donnée par :

$$f(x) = \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}$$

$$\text{avec} \quad \alpha, \beta > 0, x \in [0,1] \text{ et } B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

On a $E(X) = \alpha/(\alpha+\beta)$ et $V(X) = \alpha\beta/[(\alpha+\beta+1)(\alpha+\beta)^2]$

- Bêta4 (α, β, c, d) : la densité de cette loi est donnée par :

$$f(x) = \frac{1}{B(\alpha,\beta)} \frac{(x-c)^{\alpha-1}(d-x)^{\beta-1}}{(d-c)^{\alpha+\beta-1}}, \quad \text{avec } \alpha, \beta > 0, x \in [c,d]$$

$$c, d \in \mathbb{R} \text{ et } B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

On a $E(X) = \frac{c+(c-d)\alpha}{(\alpha+\beta)}$ et $V(X) = \frac{(c-d)^2\alpha\beta}{(\alpha+\beta+1)(\alpha+\beta)^2}$

Pour la loi Bêta de type I, la distribution est dans l'intervalle $[0, 1]$. La loi Bêta4 est obtenue par un simple changement de variable de la loi Bêta de type I de telle sorte que la distribution soit sur l'intervalle [c, d].

- Binomiale (n, p) : la densité de cette loi est donnée par :

$$P(X = x) = C_n^x p^x (1-p)^{n-x}, \quad \text{avec } n, x \in \mathbb{N}, n > 0, x \in [0, n], p \in [0, 1]$$

On a $E(X) = np$ et $V(X) = np(1-p)$

n est le nombre d'essais, et p la probabilité de succès. La loi binomiale est la loi du nombre de succès pour n essais, sachant que la probabilité de succès vaut p. La loi binomiale peut être vue comme la loi de n tirages dans une loi de Bernoulli.

- Binomiale négative (n, p) de type I : la densité de cette loi est donnée par :

$$P(X = x) = C_{n+x-1}^{x-1} p^n (1-p)^x, \quad \text{avec } n, x \in \mathbb{N}, n > 0, p \in [0, 1]$$

On a $E(X) = n(1-p)/p$ et $V(X) = n(1-p)/p^2$

n est le nombre de succès et p la probabilité de succès. La loi binomiale négative de type I est la loi du nombre de tirages x sans succès nécessaires avant d'avoir obtenus n succès.

- Binomiale négative (k, p) de type II : la densité de cette loi est donnée par :

$$P(X = x) = \frac{\Gamma(k+x)p^x}{x!\Gamma(k)(1+p)^{k+x}}, \quad \text{avec } x \in N, k, p > 0$$

On a $E(X) = kp$ et $V(X) = kp(p+1)$

La loi binomiale négative de type II permet de représenter des phénomènes discrets fortement hétérogènes. Lorsque k tend vers l'infini, la loi binomiale négative de type II tend vers une loi de Poisson de paramètre $\lambda = kp$.

- $Khi^2(df)$: la densité de cette loi est donnée par :

$$f(x) = \frac{(1/2)^{df/2}}{\Gamma(df/2)} x^{\frac{df}{2}-1} e^{-x/2}, \quad \text{avec } x > 0, df \in \mathbb{N}^*$$

On a $E(X) = df$ et $V(X) = 2df$

La loi du Khi² correspond à la loi de la somme des carrés de df lois normales centrées réduites (lois normales standard). Elle est très utilisée pour tester des hypothèses.

- Erlang (k, λ) : la densité de cette loi est donnée par :

$$f(x) = \lambda^k x^{k-1} \frac{e^{-\lambda x}}{(k-1)!}, \quad \text{avec } x \geq 0 \text{ et } k, \lambda > 0 \text{ et } k \in \mathbb{N}$$

On a $E(X) = k/\lambda$ et $V(X) = k/\lambda^2$

k est le paramètre de forme de la loi et λ est le paramètre de taux.

Cette distribution, développée par le scientifique danois A. K. Erlang (1878-1929) pour l'étude du trafic téléphonique, est utilisée de manière plus générale pour l'étude des files d'attente.

Remarque : lorsque $k = 1$, cette distribution est équivalente à la distribution exponentielle, et la loi Gamma à deux paramètres est une généralisation de la loi d'Erlang au cas où k est un réel et non un entier (par ailleurs on utilise le paramètre d'échelle $\beta = 1/\lambda$).

- Exponentielle (λ) : la densité de cette loi est donnée par :

$$f(x) = \lambda \exp(-\lambda x), \quad \text{avec } x > 0 \text{ et } \lambda > 0$$

On a $E(X) = 1/\lambda$ et $V(X) = 1/\lambda^2$

La loi exponentielle est souvent utilisée pour étudier la durée de vie en contrôle qualité.

- Fisher (df_1, df_2) : la densité de cette loi est donnée par :

$$f(x) = \frac{1}{xB(df_1/2, df_2/2)} \left(\frac{df_1 x}{df_1 x + df_2}\right)^{df_1/2} \left(1 - \frac{df_1 x}{df_1 x + df_2}\right)^{df_2/2}$$

$$\text{avec } x > 0 \text{ et } df_1, df_2 \in \mathbb{N}^*$$

On a $E(X) = df_2/(df_2 - 2)$ si $df_2 > 2$, et $V(X) = \frac{2df_2^2(df_1+df_2-2)}{df_1(df_2-2)^2(df_2-4)}$ si $df_2 > 4$

La loi de Fisher, du nom du biologiste, généticien et statisticien Ronald Aylmer Fisher (1890-1962), correspond au rapport de deux lois du Khi². Elle est très utilisée pour tester des hypothèses.

- Fisher-Tippett (β, μ) : la densité de cette loi est donnée par :

$$f(x) = \frac{1}{\beta} \exp\left(-\frac{x-\mu}{\beta} - \exp\left(-\frac{x-\mu}{\beta}\right)\right), \quad \text{avec } \beta > 0$$

On a $E(X) = \mu + \beta\gamma$ et $V(X) = (\pi\beta)^2/6$ où γ est la constante de Euler-Mascheroni.

La loi de Fisher-Tippett, aussi appelée loi Log-Weibull, ou loi généralisée des valeurs extrêmes, est utilisée dans l'étude de phénomènes extrêmes. La loi de Gumbel est un cas particulier de la loi de Fisher-Tippett avec $\beta = 1$ et $\mu = 0$.

- Gamma (k, β, μ) : la densité de cette loi est donnée par :

$$f(x) = (x-\mu)^{k-1}\ \frac{e^{-(x-\mu)/\beta}}{\beta^k \Gamma(k)}, \quad \text{avec } x > \mu \text{ et } k, \beta > 0$$

On a $E(X) = \mu + k\beta$ et $V(X) = k\beta^2$

k est le paramètre de forme de la loi et β est le paramètre d'échelle.

- GEV (β, k, μ): la densité de cette loi est donnée par :

$$f(x) = \frac{1}{\beta}\left(1 + k\frac{x-\mu}{\beta}\right)^{-1/k-1} \exp\left(-\left(1 + k\frac{x-\mu}{\beta}\right)^{-1/k}\right), \quad \text{avec} \quad \beta > 0$$

On a $E(X) = \mu + \frac{\beta}{k}\Gamma(1+k)$ et $V(X) = \left(\frac{\beta}{k}\right)^2 \left(\Gamma(1+2k) - \Gamma^2(1+k)\right)$

La loi GEV (Generalized Extreme Values) est très utilisée en hydrologie pour modéliser les phénomènes de crues. k est classiquement compris entre -0.6 et 0.6.

- Gumbel : la densité de cette loi est donnée par :

$$f(x) = \exp(-x - \exp(-x))$$

On a $E(X) = \gamma$ et $V(X) = \pi^2/6$ où γ est la constante de Euler-Mascheroni (0.5772156649…).

La loi de Gumbel, du nom de Emil Julius Gumbel (1891-1966), est un cas particulier de la loi de Fisher-Tippett avec $\beta = 1$ et $\mu = 0$. Elle est utilisée dans l'étude de phénomènes extrêmes comme les précipitations ou les crues maximales et les magnitudes maximales de tremblement de terre.

- Logistique (μ, s) : la densité de cette loi est donnée par :

$$f(x) = \frac{e^{-\frac{(x-\mu)}{s}}}{s(1 + e^{-\frac{(x-\mu)}{s}})}, \quad \text{avec } s > 0$$

On a $E(X) = \mu$ et $V(X) = (\pi s)^2/3$

- Lognormale (μ, σ) : la densité de cette loi est donnée par :

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(\ln(x)-\mu)^2}{2\sigma^2}}, \quad \text{avec } x, \sigma > 0$$

On a $E(X) = exp(\mu + \sigma^2/2)$ et $V(X) = \left(exp(\sigma^2) - 1\right) exp(2\mu + \sigma^2)$

- Lognormale2 (m, s) : la densité de cette loi est donnée par :

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(\ln(x)-\mu)^2}{2\sigma^2}}, \text{ avec } x, \sigma > 0$$

On a :

$\mu = ln(m) - ln(1 + s^2/m^2)/2$ et $\sigma^2 = ln(1 + s^2/m^2)$

Et :

$E(X) = m$ et $V(X) = s^2$

Cette distribution est simplement une reparamétrisation de la loi Lognormale.

- Normale (μ, σ) : la densité de cette loi est donnée par :

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad \text{avec } \sigma > 0$$

On a $E(X) = \mu$ et $V(X) = \sigma^2$

- Normale standard : la densité de cette loi est donnée par :

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

On a $E(X) = 0$ et $V(X) = 1$

Cette loi est un cas particulier de la loi normale, avec $\mu = 0$ et $\sigma = 1$. Elle est aussi appelée loi normale centrée réduite.

- Pareto (a, b) : la densité de cette loi est donnée par :

$$f(x) = \frac{ab^a}{x^{a+1}}, \text{ avec } a, b > 0 \text{ et } x \geq b$$

On a $E(X) = ab/(a-1)$ et $V(X) = \frac{ab^2}{((a-1)^2(a-2))}$

La loi de Pareto, du nom de l'économiste italien Vilfredo Pareto (1848-1923), est aussi connue sous le nom de loi de Bradford. Cette loi a d'abord été utilisée pour représenter la répartition des richesses dans la société, avec notamment le principe de Pareto, selon lequel 80% des richesses d'un pays sont détenus par 20% de la population.

- PERT (a, m, b) : la densité de cette loi est donnée par :

$$f(x) = \frac{1}{B(\alpha,\beta)} \frac{(x-a)^{\alpha-1}(b-x)^{\beta-1}}{(b-a)^{\alpha+\beta-1}}, \text{ avec } \alpha,\beta > 0, x \in [a,b]$$

$$a,b \in \mathbb{R} \text{ et } B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

$$\alpha = \frac{4m+b-5a}{b-a}$$

$$\beta = \frac{5b-a-4m}{b-a}$$

On a $E(X) = (b-a)\alpha/(\alpha+\beta)$ et $V(X) = (b-a)^2\alpha\beta/\left((\alpha+\beta+1)\left(\alpha+\beta\right)^2\right)$

La loi de PERT est donc un cas particulier de la loi Bêta4, définie par son intervalle de définition $[a,b]$ et sa valeur la plus probable m (le mode). PERT est l'acronyme de *Program Evaluation and Review Technique* , une méthode de gestion et de planification de projet. La méthodologie et la distribution PERT ont été utilisées pour la première fois pour le projet de développement des missiles Polaris lancés depuis des sous-marins par la marine américaine et Lockheed de 1956 à 1960 (Clark 1962). La distribution PERT permet de modéliser le temps probable nécessaire à une équipe pour terminer son projet. La loi triangulaire, plus simple, permet aussi de modéliser ce type de phénomènes avec les trois mêmes paramètres.

- Poisson (λ): la densité de cette loi est donnée par :

$$P(X=x) = \frac{\exp(-\lambda)\lambda^x}{x!}, \quad \text{avec } x \in \mathbb{N} \text{ et } \lambda > 0$$

On a $E(X) = \lambda$ et $V(X) = \lambda$

La loi de Poisson, découverte par le mathématicien et astronome Siméon-Denis Poisson (1781-1840) qui fut élève de Laplace, Lagrange et Legendre, est souvent utilisée pour étudier des phénomènes de file d'attente.

- Student (df) : la densité de cette loi est donnée par :

$$f(x) = \frac{\Gamma((df+1/2))}{\sqrt{\pi df}\Gamma(df/2)}(1+x^2/df)^{-(df+1)/2}, \text{ avec } df > 0$$

On a $E(X) = 0$ si $df > 1$ et $V(X) = df/(df-2)$ si $df > 2$

La loi de Student, du nom que se donnait le chimiste et statisticien anglais William Sealy Gosset (1876-1937) afin de préserver son anonymat (la brasserie Guinness interdisait à ses employés de publier, suite à la publication par un autre chercheur d'informations confidentielles) est la loi de la moyenne de df variables distribuées suivant une loi normale centrée réduite. Lorsque $df = 1$, la loi de Student est une loi de Cauchy dont la particularité est de n'avoir ni espérance ni variance.

- Trapézoïdale (a, b, c, d) : la densité de cette loi est donnée par :

$$\begin{cases} f(x) = \frac{2(x-a)}{(d+c-b-a)(b-a)}, & x \in [a, b] \\ f(x) = \frac{2}{(d+c-b-a)}, & x \in [b, c] \\ f(x) = \frac{2(d-x)}{(d+c-b-a)(d-c)}, & x \in [c, d] \\ f(x) = 0, & x < a, \quad x > d \\ \text{avec } a < b < c < d \end{cases}$$

On a $E(X) = \frac{d^2+c^2-b^2-a^2+cd-ab}{3(d+c-b-a)}$ et $V(X) = \frac{(c+d)(c^2+d^2)-(a+b)(a^2+b^2)}{6(d+c-b-a)} - E^2(X)$

Cette loi est utile pour représenter un phénomène dont on sait qu'il peut prendre des valeurs entre deux extrêmes, mais pour lequel un intervalle plus restreint paraît plus raisonnable.

- Triangulaire (a, m, b) : la densité de cette loi est donnée par :

$$\begin{cases} f(x) = \frac{2(x-a)}{(b-a)(m-a)}, & x \in [a, m] \\ f(x) = \frac{2(b-x)}{(b-a)(b-m)}, & x \in [m, b] \\ f(x) = 0, & x < a, \quad x < b \\ \text{avec } a < m < b \end{cases}$$

On a $E(X) = (a + m + b)/3$ et $V(X) = (a^2 + m^2 + b^2 - ab - am - bm)/18$

- TriangulaireQ (q_1, m, q_2, p_1, p_2): cette loi est une reparamétrisation de la loi triangulaire. Une première étape nécessite l'estimation des paramètres a et b de la distribution triangulaire pour savoir à quels quantiles q_1 et q_2 correspondent les pourcentages p_1 et p_2. Une fois ceci fait, on peut utiliser la fonction de densité ou de répartition triangulaire.

- Uniforme (a, b) : la densité de cette loi est donnée par :

$$f(x) = \frac{1}{b-a}, \text{ avec } b > a \text{ et } x \in [a, b]$$

On a $E(X) = (a + b)/2$ et $V(X) = (b - a)^2/12$

La loi uniforme (0, 1) est très utilisée pour les simulations. Comme la fonction de répartition de toutes les lois est comprise entre 0 et 1, un échantillon tiré dans une loi Uniforme (0,1) permet

d'obtenir un échantillon dans toutes les lois dont on sait calculer l'inverse.

- Uniforme discrète (a, b) : la densité de cette loi est donnée par :

$$P[X = x] = \frac{1}{b - a + 1}, \text{ avec } (a, b, x) \in \mathbb{N}^3, x \in [a, b]$$

On a $E(X) = (a + b)/2$ et $V(X) = [(b - a + 1)^2 - 1]/12$

La loi uniforme discrète correspond au cas particulier où la loi uniforme est restreinte à des nombre entiers.

- Weibull (β) : la densité de cette loi est donnée par :

$$f(x) = \beta x^{\beta-1} \exp(-x^\beta), \text{ avec } x > 0 \text{ et } \beta > 0$$

On a $E(X) = \Gamma(\frac{1}{\beta} + 1)$ et $V(X) = \Gamma(\frac{2}{\beta} + 1) - \Gamma^2(\frac{1}{\beta} + 1)$

Le paramètre β est le paramètre de forme de la loi de Weibull.

- Weibull (β, γ) : la densité de cette loi est donnée par :

$$f(x) = \frac{\beta}{\gamma}(\frac{x}{\gamma})^{\beta-1} e^{-(\frac{x}{\gamma})^\beta}, \text{ avec } x > 0, \text{ et } \beta, \gamma > 0$$

On a $E(X) = \gamma\Gamma\left(\frac{1}{\beta} + 1\right)$ et $V(X) = \gamma^2\left[\Gamma\left(\frac{2}{\beta} + 1\right) - \Gamma^2\left(\frac{1}{\beta} + 1\right)\right]$

Le paramètre β est le paramètre de forme et le paramètre γ est le paramètre d'échelle. Lorsque $\beta = 1$, la loi de Weibull est une loi exponentielle de paramètre $1/\gamma$.

- Weibull (β, γ, μ): la densité de cette loi est donnée par :

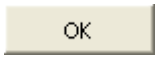
$$f(x) = \frac{\beta}{\gamma}\left(\frac{x - \mu}{\gamma}\right)^{\beta-1} e^{-\left(\frac{x-\mu}{\gamma}\right)^\beta}, \text{ avec } x > \mu, \text{ et } \beta, \gamma > 0$$

On a $E(X) = \mu + \gamma\Gamma\left(\frac{1}{\beta} + 1\right)$ et $V(X) = \gamma^2\left[\Gamma\left(\frac{2}{\beta} + 1\right) - \Gamma^2\left(\frac{1}{\beta} + 1\right)\right]$

La loi de Weibull, du nom du suédois Ernst Hjalmar Waloddi Weibull (1887-1979), est très utilisée en contrôle qualité et en analyse de survie. Le paramètre β est le paramètre de forme et le paramètre γ est le paramètre d'échelle. Lorsque $\beta = 1$ et $\mu = 0$, la loi de Weibull est une loi exponentielle de paramètre $1/\gamma$.

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.

Onglet **Général**:

Données : sélectionnez des données quantitatives. Si plusieurs échantillons sont sélectionnés, XLSTAT fera les calculs pour chacun des échantillons indépendamment, tout en vous permettant de superposer les histogrammes si vous le souhaitez (voir l'onglet Graphiques). Si des en-têtes ont été sélectionnés, veuillez vérifier que l'option « Libellés des échantillons » est activée.

Type de données :

Continues : choisissez cette option pour que XLSTAT considère que vos données sont continues.

Discrètes : choisissez cette option pour que XLSTAT considère que vos données sont discrètes.

Sous-échantillons : activez cette option puis sélectionnez une colonne (mode colonnes) ou une ligne (mode lignes) contenant les descripteurs d'échantillons. L'utilisation de cette option permet d'obtenir un histogramme par sous-échantillon et donc de comparer la distribution des données entre les sous-échantillons. Si un en-tête a été sélectionné, veuillez vérifier que l'option « Libellés des échantillons » est activée.

- Libellés Variable-Modalité : activez cette option pour utiliser des libellés longs pour l'affichage des résultats. Les libellés Variable-Modalité sont composés du nom de la variable comme préfixe, et de la modalité du sous-échantillon comme suffixe.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des échantillons : activez cette option si la première ligne des données sélectionnées (données, sous échantillons, poids) contient un libellé.

Poids : activez cette option si vous voulez pondérer les observations. Si vous n'activez pas cette option, les poids seront tous considérés comme valant 1. Les poids doivent être impérativement supérieurs ou égaux à 0. Si un en- tête de colonne a été sélectionné, veuillez vérifier que l'option « Libellés des échantillons » est activée.

Onglet **Options**:

Intervalles : choisissez l'une des options suivantes pour définir les intervalles de l'histogramme :

- **Nombre** : choisissez cette option pour entrer le nombre d'intervalles à créer.
- **Amplitude** : choisissez cette option pour définir une amplitude fixe pour les intervalles.
- **Définis par l'utilisateur** : sélectionnez une colonne contenant en ordre croissant la borne inférieure du premier intervalle, et la borne supérieure de tous les intervalles.
- **Minimum** : activez cette option pour entrer la valeur de la borne inférieure du premier intervalle. Cette valeur doit être inférieure ou égale au minimum de la série.

Comparer les sous-échantillons : cette option n'est active que si une colonne de sous-échantillons a été sélectionnée. Activez cette option pour afficher les différents sous-échantillons sur un même histogramme.

- **Comparer à l'échantillon total** : activez cette option pour que les statistiques descriptives et les graphiques soient aussi affichés pour l'échantillon total.

Onglet **Données manquantes** :

Ne pas accepter les valeurs manquantes : activez cette option pour que XLSTAT empêche la poursuite des calculs si des valeurs manquantes sont détectées.

Supprimer les observations :

- **Pour l'échantillon correspondant** : activez cette option pour ne pas prendre en compte une observation dont l'une des données est manquante, uniquement pour les échantillons pour lesquels une donnée est manquante.

- **Pour tous les échantillons** : activez cette option pour ne pas prendre en compte une observation dont l'une des données est manquante, pour tous les échantillons sélectionnés.

Estimer les données manquantes : activez cette option pour estimer les données manquantes en utilisant la moyenne de l'échantillon.

Onglet **Sorties**:

Statistiques descriptives : activez cette option pour afficher les statistiques descriptives des échantillons.

Onglet **Graphiques** :

Histogrammes : activez cette option pour afficher les histogrammes des échantillons. Pour la distribution théorique, la fonction de densité est affichée.

- **Barres** : choisissez cette option pour afficher des histogrammes avec une barre pour chaque intervalle.
- **Lignes continues** : choisissez cette option pour afficher des histogrammes avec une ligne continue.

Histogrammes cumulés : activez cette option pour afficher les histogrammes cumulés des échantillons.

- **Basés sur l'histogramme** : choisissez cette option pour afficher des histogrammes cumulés basés sur la même définition d'intervalles que les histogrammes.
- **Fonction de répartition empirique** : choisissez cette option pour afficher des histogrammes cumulés qui correspondent en réalité à la fonction de répartition empirique de l'échantillon.

Ordonnées des histogrammes : choisissez quelle grandeur doit être utilisée pour les histogrammes : densité, effectif ou fréquence.

Afficher une distribution : activez cette option pour comparer les histogrammes des échantillons sélectionnés à une fonction de densité et/ou pour comparer les histogrammes des échantillons sélectionnés à une fonction de répartition. Choisissez alors la loi à utiliser, puis, si nécessaire, entrez la valeur de ses paramètres. L'option **automatique** permet de laisser XLSTAT identifier la distribution s'ajuste le mieux (déterminé sur la base d'un test de Kolmogorov-Smirnov).

Résultats

Statistiques simples : dans ce tableau sont affichées pour tous les échantillons les statistiques descriptives suivantes : le nombre d'observations, le nombre de données manquantes, le nombre de données non manquantes, la moyenne, et l'écart-type (non biaisé).

Histogrammes : les histogrammes sont affichés. Si vous le souhaitez, vous pouvez modifier la couleur des lignes, les échelles, et les titres comme avec n'importe quel graphique Excel.

Statistiques descriptives pour les intervalles : dans ce tableau sont affichés pour chaque intervalle sa borne inférieure, sa borne supérieure, le nombre de valeurs de l'échantillon étant comprises dans l'intervalle (effectif), la fréquence (l'effectif divisé par l'effectif total de l'échantillon), et la densité (le rapport de la fréquence sur la taille de l'intervalle).

Exemple

Un exemple de génération d'histogramme est disponible sur le Centre d'aide XLSTAT à l'adresse suivante :

http://www.xlstat.com/demo-histof.htm

Bibliographie

Chambers J.M., Cleveland W.S., Kleiner B. and Tukey P.A. (1983). Graphical Methods for Data Analysis. Duxbury, Boston.

Jacoby W. G. (1997). Statistical Graphics for Univariate and Bivariate Data. Sage Publications, London.

Wilkinson L. (1999). The Grammar of Graphics, Springer Verlag, New York.

Estimation de densité par noyau

Utiliser cet outil pour calculer et afficher une densité estimée par la méthode des noyaux à partir d'un échantillon univarié de données quantitatives. Il s'agit d'une alternative à la méthode d'affichage des histogrammes, et d'une alternative non paramétrique à l'ajustement de distribution paramétrique si vous avez besoin de réutiliser les estimations de densité.

Dans cette section :

Description

L'estimation de la densité par noyau (en anglais, *Kernel density estimation* ou *KDE*) permet d'estimer la densité d'un échantillon univarié. C'est l'une des alternatives non-paramétriques à l'ajustement paramétrique de la distribution. Alors que l'approche paramétrique nécessite la connaissance de la distribution F de la variable, et la connaissance ou l'estimation des paramètres de F, nous n'aurons besoin ici que de spécifier un noyau et une bande passante (*bandwidth*).

Soit X une variable aléatoire et $\{x_i\}$, $(i = 1, ..., n)$, un échantillon de taille n, et w_i les poids associés aux observations (ils valent 1 s'il n'y a pas de pondération particulière). Soit $W = \sum_{i=1}^{n} w_i$. L'estimation par noyau de la fonction de densité de probabilité f de X est donnée par : $$ {\widehat{f}}*{h}(x)={\frac {1}{Wh}}\sum _{i=1}^{n}{w_i}K{\Big (}{\frac {x-x}{h}}{\Big)} $$

K est la fonction noyau. h est la bande passante. Les fonctions noyau sont des fonctions normées, intégrables et symétriques. Il est également important de noter que si K est un noyau, alors $\lambda K(\lambda z)$ est également un noyau, ce qui signifie que l'on peut changer la largeur de bande h tout en gardant les propriétés mathématiques inchangées.

Bien qu'elle ne soit pas aussi importante que la bande passante, la fonction noyau influence la densité. Les noyaux suivants sont disponibles dans XLSTAT :

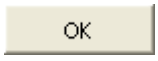
- Biweight (ou Quartic): $$ \begin{cases} K(z)=\frac{15}{16}(1-z^2)2 &|z| \leq 1 \ K(z)=0 &|z| \gt 1 \end{cases} $$
- Cosinus $$ \begin{cases} K(z)=\frac{1}{2}(1+\cos(z\pi)) &|z| \leq 1 \ K(z)=0 &|z| \gt 1 \end{cases} $$
- Epanechnikov $$ \begin{cases} K(z)=\frac{3}{4}(1- \frac{z^2}{\sqrt{5}})/\sqrt{5} &|z| \leq \sqrt{5} \ K(z)=0 &|z| \gt \sqrt{5} \end{cases} $$

- Epanechnikov (0.75) $$ \begin{cases} K(z)=\frac{3}{4}(1-z^2) &|z| \leq 1 \ K(z)=0 &|z| \gt 1 \end{cases} $$
- Gaussien $$ \begin{cases} K(z)={\frac {1}{\sqrt {2\pi }}}e^{-{\frac {1}{2}}z^{2}} & z \in \left]-\infty, +\infty\right[\ \end{cases} $$
- Optcosine $$ \begin{cases} K(z)=\frac{\pi}{4} \cos(\frac{\pi}{2} z) &|z| \leq 1 \ K(z)=0 &|z| \gt 1 \end{cases} $$
- Parzen $$ \begin{cases} K(z)=\frac {4}{3} -8 z^{2+8|z|}3 &|z| \leq 0.5 \ K(z)=\frac {8}{3}(1-|z|)^3 & 0.5 \leq |z| \leq 1 \ K(z)=0 &|z| \gt 1 \end{cases} $$
- Triangulaire $$ \begin{cases} K(z)=(1-|z|) &|z| \leq 1 \ K(z)=0 &|z| \gt 1 \end{cases} $$
- Tricube $$ \begin{cases} K(z)=\frac{70}{81}(1-|z|^{3)}3 &|z| \leq 1 \ K(z)=0 &|z| \gt 1 \end{cases} $$
- Triweight $$ \begin{cases} K(z)=\frac{35}{32}(1-z^{2)}3 &|z| \leq 1 \ K(z)=0 &|z| \gt 1 \end{cases} $$
- Uniforme $$ \begin{cases} K(z)=0.5 &|z| \leq 0.5 \ K(z)=0 &|z| \gt 0.5 \end{cases} $$

Comme pour la largeur des intervalles des histogrammes, la bande passante h conditionne fortement la forme de la fonction de densité, y compris le nombre de modes. XLSTAT propose les options suivantes pour la largeur de bande : * Définie par l'utilisateur : Vous pouvez entrer la valeur de votre choix * Silverman(1) : $h = 0.9 * min(\hat{\sigma}, IQR/1.34)/n^{1/5}$, où IQR est l'écart interquartile non remis à l'échelle (voir Silverman, page 48) * Silverman(2) : $h = 1.06 * \hat{\sigma}/n^{1/5}$, (voir Silverman, page 45)

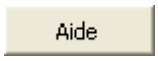
Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.



: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.


: cliquez sur ce bouton pour changer la façon dont vous chargez les données dans XLSTAT. Si le bouton possède une icône de souris, XLSTAT vous permet de faire une sélection à la souris des données. Si le bouton possède une icône de liste, XLSTAT vous permet de sélectionner les données à partir d'une liste. Si le bouton possède une icône orange, des

boutons complémentaires avec un point d'interrogation  sont affichés pour vous permettre d'importer les données à partir de fichiers.

Onglet **Général**:

Données : sélectionnez des données quantitatives. Si plusieurs échantillons sont sélectionnés, XLSTAT fera les calculs pour chacun des échantillons indépendamment. Si des en-têtes ont été sélectionnés, veuillez vérifier que l'option « Libellés des échantillons » est activée.

Poids : activez cette option si vous voulez pondérer les observations. Si vous n'activez pas cette option, les poids seront tous considérés comme valant 1. Les poids doivent être impérativement supérieurs ou égaux à 0. Si un en- tête de colonne a été sélectionné, veuillez vérifier que l'option « Libellés des échantillons » est activée.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des échantillons : activez cette option si la première ligne des données sélectionnées (données, sous échantillons, poids) contient un libellé.

Onglet **Données manquantes** :

Ne pas accepter les valeurs manquantes : activez cette option pour que XLSTAT empêche la poursuite des calculs si des valeurs manquantes sont détectées.

Supprimer les observations :

- **Pour l'échantillon correspondant** : activez cette option pour ne pas prendre en compte une observation dont l'une des données est manquante, uniquement pour les échantillons pour lesquels une donnée est manquante.
- **Pour tous les échantillons** : activez cette option pour ne pas prendre en compte une observation dont l'une des données est manquante, pour tous les échantillons sélectionnés.

Estimer les données manquantes : activez cette option pour estimer les données manquantes en utilisant la moyenne de l'échantillon.

Onglet **Sorties**:

Statistiques descriptives : activez cette option pour afficher les statistiques descriptives des échantillons.

Estimateurs de la densité (noyau) : activez cette option pour afficher les estimations de la densité pour chaque observation.

Onglet **Graphiques** :

Densité de noyau : activez cette option pour afficher un histogramme derrière la courbe de densité de noyau. Utilisez les options suivantes pour contrôler l'affichage de l'histogramme :

Histogrammes : activez cette option pour afficher les histogrammes des échantillons.

- **Nombre de points** : Définissez le nombre de points pour lesquels la densité est calculée. Si le minimum et le maximum ne sont pas spécifiés (voir ci-dessous), les calculs sont effectués dans l'intervalle $]min - 0,1 \times amplitude, max + 0,1 \times amplitude[$.
- **Minimum** : Activez cette option pour saisir la valeur inférieure pour laquelle la courbe de densité est calculée. **Maximum** : Activez cette option pour saisir la valeur supérieure pour laquelle la courbe de densité est calculée.

Histogrammes : activez cette option pour afficher les histogrammes des échantillons.

- **Intervalles** : choisissez l'une des options suivantes pour définir les intervalles de l'histogramme :
  - **Nombre** : choisissez cette option pour entrer le nombre d'intervalles à créer.
  - **Amplitude** : choisissez cette option pour définir une amplitude fixe pour les intervalles.
  - **Définis par l'utilisateur** : sélectionnez une colonne contenant en ordre croissant la borne inférieure du premier intervalle, et la borne supérieure de tous les intervalles.
  - **Minimum** : activez cette option pour entrer la valeur de la borne inférieure du premier intervalle. Cette valeur doit être inférieure ou égale au minimum de la série.

Résultats

Statistiques simples : dans ce tableau sont affichées pour tous les échantillons les statistiques descriptives suivantes : le nombre d'observations, le nombre de données manquantes, le nombre de données non manquantes, la moyenne, et l'écart-type (non biaisé).

Les résultats suivants sont affichés pour chaque échantillon :

Bande passante : XLSTAT affiche la bande passante qui a été utilisée pour les calculs. Si vous voulez modifier le tracé de la densité de noyau, vous pouvez changer la valeur de la bande passante en fonction de la valeur fournie : augmentez/diminuez la valeur de la bande passante si vous voulez une courbe plus lisse/plus détaillée.

Estimateurs de la densité (noyau) : dans ce tableau sont affichés les estimateurs de la densité par noyau pour chaque observation.

Graphiques : XLSTAT affiche les courbes de densité par noyau. Si l'option correspondante a été activée, un histogramme est également affiché. Si vous le souhaitez, vous pouvez modifier

la couleur des lignes, des échelles, des titres comme pour tout graphique Excel.

Statistiques descriptives pour les intervalles : dans ce tableau sont affichés pour chaque intervalle sa borne inférieure, sa borne supérieure, le nombre de valeurs de l'échantillon étant comprises dans l'intervalle (effectif), la fréquence (l'effectif divisé par l'effectif total de l'échantillon), et la densité (le rapport de la fréquence sur la taille de l'intervalle).

Exemple

Un exemple d'utilisation de cet outil est disponible sur le Centre d'aide XLSTAT à l'adresse suivante :

http://www.xlstat.com/demo-kdff.htm

Bibliographie

Chambers J.M., Cleveland W.S., Kleiner B. and Tukey P.A. (1983). Graphical Methods for Data Analysis. Duxbury, Boston.

Parzen E. (1962). On estimation of a probability density function and mode. *Annals of Mathematical Statistics*, **33**, 1065-1076.

Silverman B. W. (1986). Density Estimation for Statistics and Data Analysis. Chapman & Hal, London.

Tests de normalité

Utilisez cet outil pour vérifier si un échantillon peut être considéré comme étant distribué suivant une loi normale. L'outil ajustement d'une loi de probabilité permet d'estimer les paramètres de la loi normale mais les tests qui sont proposés ne sont pas aussi bien adaptés que ceux proposés ici.

Dans cette section :

Description

Supposer la normalité d'un échantillon ou d'une statistique est commun en statistique. Pourtant, la vérification de l'hypothèse de normalité est souvent négligée. Par exemple, la normalité des résidus obtenus lors d'une régression linéaire est rarement testée, alors qu'elle conditionne la qualité des intervalles de confiance autour des paramètres et des prédictions.

XLSTAT propose quatre tests pour tester la normalité d'un échantillon :

le test de Shapiro-Wilk bien adapté aux échantillons de moins de 5000 observations ;

le test d'Anderson-Darling proposé par Stephens (1974) est une modification du test de Kolmogorov-Smirnov adaptée à plusieurs lois dont la loi normale, pour le cas où les paramètres de la loi ne sont pas connus et doivent donc être estimés ;

le test de Lilliefors est une modifiication du test de Kolmogorov-Smirnov adapté au cas de la normalité dans le cas où les paramètres de la loi, la moyenne et la variance, ne sont pas connus et doivent donc être estimés ;

le test de Jarque-Bera qui est d'autant plus performant que le nombre de données est important.

Afin de vérifier visuellement si un échantillon suit une loi normale, il est possible d'utiliser les graphiques P-P et les graphiques Q-Q :

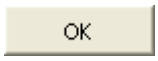
Graphiques P-P (loi normale) : les graphiques Probabilité-Probabilité (*P-P plots* en anglais) permettent de comparer la fonction de répartition empirique d'un échantillon à celle d'un échantillon distribué suivant une loi normale de même moyenne et même variance. Si

l'échantillon suit une loi normale, les points doivent être confondus avec la première bissectrice du plan.

Graphiques Q-Q (loi normale) : les graphiques Quantile-Quantile (*Q-Q plots* en anglais) permettent de comparer les quantiles de l'échantillon à ceux d'un échantillon distribué suivant une loi normale de même moyenne et même variance. Si l'échantillon suit une loi normale, les points doivent être confondus avec la première bissectrice du plan.

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.


: cliquez sur ce bouton pour changer la façon dont vous chargez les données dans XLSTAT. Si le bouton possède une icône de souris, XLSTAT vous permet de faire une sélection à la souris des données. Si le bouton possède une icône de liste, XLSTAT vous permet de sélectionner les données à partir d'une liste. Si le bouton possède une icône orange, des boutons complémentaires avec un point d'interrogation  sont affichés pour vous permettre d'importer les données à partir de fichiers.

Onglet **Général**:

Données : sélectionnez des données quantitatives. Si plusieurs échantillons sont sélectionnés, XLSTAT testera la normalité pour chacun des échantillons indépendamment. Si des en-têtes ont été sélectionnés, veuillez vérifier que l'option « Libellés des échantillons » est activée.

Poids : activez cette option si vous voulez pondérer les observations. Si vous n'activez pas cette option, les poids seront tous considérés comme valant 1. Les poids doivent être

impérativement supérieurs ou égaux à 0. Si un en- tête de colonne a été sélectionné, veuillez vérifier que l'option « Libellés des échantillons » est activée.

Test de Shapiro-Wilk : activez cette option pour effectuer un test de Shapiro-Wilk.

Test d'Anderson-Darling : activez cette option pour effectuer un test d'Anderson-Darling.

Test de Lilliefors : activez cette option pour effectuer un test de Lilliefors.

Test de Jarque-Bera : activez cette option pour effectuer un test de Jarque-Bera.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des échantillons : activez cette option si la première ligne des données sélectionnées (données, sous échantillons, poids) contient un libellé.

Niveau de signification (%) : entrez le niveau de signification pour les tests.

Sous-échantillons : activez cette option puis sélectionnez une colonne (mode colonnes) ou une ligne (mode lignes) contenant les descripteurs d'échantillons. L'utilisation de cette option permet de calculer les tests de normalité pour chacun des sous-échantillons. Si un en-tête a été sélectionné, veuillez vérifier que l'option « Libellés des échantillons » est activée.

Onglet **Données manquantes** :

Supprimer les observations :

- **Pour l'échantillon correspondant** : activez cette option pour ne pas prendre en compte une observation dont l'une des données est manquante, uniquement pour les échantillons pour lesquels une donnée est manquante.
- **Pour tous les échantillons** : activez cette option pour ne pas prendre en compte une observation dont l'une des données est manquante, pour tous les échantillons sélectionnés.

Estimer les données manquantes : activez cette option pour estimer les données manquantes en utilisant la moyenne de l'échantillon.

Onglet **Sorties**:

Statistiques descriptives : activez cette option pour afficher les statistiques descriptives des échantillons.

Onglet **Graphiques** :

Graphiques P-P : activez cette option pour afficher les graphiques probabilité-probabilité basés sur la loi normale.

Graphiques Q-Q : activez cette option pour afficher les graphiques quantile- quantile basés sur la loi normale.

Résultats

Pour chaque test demandé sont affichées les statistiques relatives au test, dont notamment la p-value qui est ensuite utilisée pour l'interprétation du test par comparaison avec le seuil de signification choisi.

S'ils ont été demandés, les P-P et Q-Q plots sont ensuite affichés.

Exemple

Un exemple de test de normalité est disponible sur le Centre d'aide XLSTAT à l'adresse suivante :

http://www.xlstat.com/demo-normf.htm

Bibliographie

Anderson T.W. and Darling D.A. (1952). Asymptotic theory of certain "Goodness of Fit" criteria based on stochastic processes. *Annals of Mathematical Statistic*, **23**, 193-212.

Anderson T.W. and Darling D.A. (1954). A test of goodness of fit. *Journal of the American Statistical Association*, **49**, 765-769.

D'Agostino R.B. and Stephens M.A. (1986). Goodness-of-fit techniques. Marcel Dekker, New York.

Dallal G.E. and Wilkinson L. (1986). An analytic approximation to the distribution of Lilliefors's test statistic for normality. *Statistical Computing*, **40**, 294-296.

Jarque C.M. and Bera A.K. (1980). Efficient tests for normality, heteroscedasticity and serial independence of regression residuals. *Economic Letters*, **6**, 255-259.

Lilliefors H. (1967). On the Kolmogorov-Smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association*, **62**, 399-402.

Royston P. (1982). An extension of Shapiro and Wilks' W test for normality to large samples. *Applied Statistics*, **31**, 115-124.

Royston P. (1982). Algorithm AS 181: the W test for normality. *Applied Statistics*, **31**, 176-180.

Royston P. (1995). A remark on Algorithm AS 181: the W test for normality. *Applied Statistics*, **44**, 547-551.

Stephens M. A. (1974). EDF statistics for goodness of fit and some comparisons. *Journal of the American Statistical Association*, **69**, 730-737.

Stephens M. A. (1976). Asymptotic results for goodness-of-fit statistics with unknown parameters. *Annals of Statistics*, **4**, 357-369.

Shapiro S. S. and Wilk M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, **52**, 3 and 4, 591-611.

Thode H.C. (2002). Testing for normality. Marcel Dekker, New York, USA.

Rééchantillonnage

Utilisez cet outil pour calculer des statistiques descriptives par rééchantillonnage pour un ensemble de variables quantitatives.

Dans cette section :

Description

Avec l'amélioration des capacités de calcul, le rééchantillonnage a connu un essor important ces dernières années. Cette technique permet d'obtenir des statistiques sans formuler d'hypothèse quant à la distribution. Le principe est simple : à partir d'un échantillon, on obtient par tirage aléatoire un autre échantillon qui nous permettra de recalculer les statistiques qui nous intéressent. En répétant cette étape un grand nombre de fois, on pourra obtenir une distribution empirique de la statistique étudiée. Ainsi, on obtient l'écart-type et un intervalle de confiance associé à n'importe quelle statistique en utilisant du rééchantillonnage sans aucune hypothèse de distribution.

XLSTAT propose d'appliquer ces méthodes sur un certain nombre de statistiques descriptives classiques.

Plusieurs méthodes de rééchantillonnage existent et sont disponibles dans XLSTAT :

- **Le Bootstrap** : c'est l'approche la plus connue et la plus utilisée. Elle est basée sur les travaux d'Efron et Tibisharni (1993). Le rééchantillonnage bootstrap est basé sur des tirages avec remise des observations de l'échantillon original. Les échantillons obtenus ont n observations. Il faut spécifier le nombre de répétitions (fixé par défaut à 100).

- **Le tirage aléatoire** : on tire aléatoirement un certain nombre d'observations sans remises et on répète cette opération. Il faut alors définir le nombre d'observations à tirer et le nombre de tirages.

- **Le Jackknife** : la procédure de rééchantillonnage est basée sur la suppression d'une observation de l'échantillon original (de taille n). Chaque sous-échantillon a alors $n-1$ observations, on répète alors n fois cette procédure. Cette procédure est moins robuste que la première.

XLSTAT vous propose un nombre important de statistiques descriptives et de graphiques qui vous permettront d'avoir un premier aperçu de la distribution associée à ces statistiques.

Bien que vous puissiez sélectionner plusieurs variables (ou échantillons) à la fois, XLSTAT calcule l'ensemble des statistiques descriptives pour chacun des échantillons indépendamment.

Statistiques descriptives :

Soit un échantillon composé de n données quantitatives $\{x_1, x_2, \ldots, x_n\}$, dont les poids respectifs sont $\{w_1, w_2, \ldots, w_n\}$.

- **Somme** : la somme pondérée des données est définie par :

$$S = \sum_{i=1}^{n} w_i x_i$$

- **Moyenne** : la moyenne de l'échantillon est définie par :

$$\mu = \frac{S}{S_W}$$

- **Variance n** : la variance de l'échantillon est définie par :

$$s(n)^2 = \frac{\sum_{i=1}^{n} w_i (x_i - \mu)^2}{S_W}$$

Remarque 1 : lorsque tous les poids valent 1, la variance est la somme des écarts quadratiques à la moyenne, divisée par n, d'où la dénomination.

Remarque 2 : la variance n est une estimation biaisée de la variance, qui suppose que l'échantillon est bien représentatif de la population totale. La variance $n-1$ est calculée au contraire en tenant compte d'une approximation liée à l'échantillonnage.

- **Variance n-1** : la variance estimée de l'échantillon est définie par :

$$s(n-1)^2 = \frac{\sum_{i=1}^{n} w_i (x_i - \mu)^2}{S_W - S_W / n}$$

Remarque 1 : lorsque tous les poids valent 1, la variance est la somme des écarts quadratiques à la moyenne, divisée par $n-1$, d'où la dénomination.

Remarque 2 : la variance n est une estimation biaisée de la variance, qui suppose que l'échantillon est bien représentatif de la population totale. La variance $n-1$ est calculée au contraire en tenant compte d'une approximation liée à l'échantillonnage.

- **Ecart-type n** * : l'écart-type de l'échantillon défini par : $s(n)$

- **Ecart-type n-1** * : l'écart-type estimé de l'échantillon défini par : $s(n-1)$
- **Médiane** * : la médiane $Q2$ est telle que 50% des données lui sont inférieur.
- **1er quartile** * : le premier quartile $Q1$ est défini comme la valeur telle que 25% des données lui sont inférieur.
- **3èmequartile** * : le troisième quartile $Q3$ est défini comme la valeur telle que 75% des données lui sont inférieur.
- **Coefficient de variation** * : ce coefficient n'est calculé que si la moyenne de l'échantillon n'est pas nulle. Il est défini par $CV = s(n)/\mu$. Ce coefficient mesure la dispersion d'un échantillon relativement à sa moyenne. Il permet de comparer la dispersion d'échantillons dont les échelles ou les moyennes sont sensiblement différentes.
- **Ecart-type de la moyenne** * : cette statistique est définie par :

$$s_\mu = \frac{s_n}{\sqrt{n-1}}$$

- **Ecart absolu moyen** * : comme l'écart-type ou la variance, ce coefficient mesure la dispersion (ou variabilité) de l'échantillon. Il est défini par :

$$e(\mu) = \frac{\sum_{i=1}^{n} w_i |x_i - \mu|}{S_W}$$

- **Ecart absolu médian** * : cette statistique correspond à la médiane des écarts absolus à la médiane.
- **Moyenne géométrique** * : cette statistique n'est calculée que si toutes les données sont strictement positives. Elle est définie par :

$$\mu_G = exp\left(\frac{1}{S_W}\sum_{i=1}^{n} w_i Ln(x_i)\right)$$

Si tous les poids sont égaux à 1, on a

$$\mu_G = \sqrt[n]{\prod_{i=1}^{n} x_i}$$

- **Ecart-type géométrique** * : cette statistique est définie par :

$$\sigma_G = \exp\left(\frac{1}{S_W}\sum_{i=1}^{n} w_i \left(\mathrm{Ln}\left(x_i\right) - \mathrm{Ln}\left(\mu_G\right)\right)^2\right)$$

- **Moyenne harmonique** * : cette statistique est définie par :

$$\mu_H = \frac{S_W}{\sum_{i=1}^{n} \frac{w_i}{x_i}}$$

Les statistiques suivies d'un astérisque (*) tiennent compte du poids des observations.

Statistiques obtenues par rééchantillonnage :

Soit une statistique S, calculée B fois sur chaque échantillon rééchantillonné, nous avons pour B rééchantillonnages dans le cas du bootstrap et du tirage aléatoire :

Moyenne : cette statistique est obtenue en faisant la moyenne sur l'ensemble des B rééchantillonnages :

$$\hat{\mu}^{\star}(S) = \frac{\sum_{i=1}^{B} \hat{S}_i}{B}$$

où S_i est la valeur de la statistique S pour l'échantillon i.

Ecart-type :

$$\hat{\sigma}^{*}(S) = \sqrt{\frac{\sum_{i=1}^{B} \left(\hat{S}_i - \hat{\mu}^{*}(S)\right)^2}{B-1}}$$

Intervalle de confiance standard : Il est donné par :

$$[S \pm u_{1-\alpha/2}\hat{\sigma}^{*}(S)]$$

Avec u est le percentile $1-\alpha/2$ de la distribution normale réduite et $1-\alpha$ est le degré de confiance retenu. Ce type d'intervalle dépend de la distribution normale et nécessite donc une hypothèse paramétrique.

Intervalle de confiance percentile : Les limites de l'intervalle de confiance sont données par les percentiles $\alpha/2$ et $1-\alpha/2$ de la distribution d'échantillonnage empirique, c'est-à-dire de la distribution des statistiques S_i.

Intervalle de confiance percentile avec correction pour le biais : Les limites de cet intervalle sont aussi obtenues en se basant sur les percentiles, avec une légère différence. On détermine d'abord la proportion p de valeurs S_i inférieures à S qui est l'estimation de la statistique sur l'échantillon original. On calcule le percentile U_p relatif à la distribution normale réduite. Les limites de l'intervalle de confiance sont les percentiles S_{a_1} et S_{a_2} de la distribution empirique des S_i. On a : $a_1 = \Phi(2u_p + u_{\alpha/2})$ et $a_2 = \Phi(2u_p + u_{1-\alpha/2})$. Pour plus de details sur cet intervalle, on peut voir Efron et Tibshirani (1993).

Cas particulier du Jackknife :

- **Moyenne** :

$$\hat{\mu}(S) = \frac{\sum_{i=1}^{n} \hat{S}_{(-i)}}{n}$$

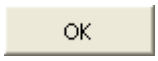
avec S(-i), statistique obtenue sur l'échantillon dans lequel l'observation I a été supprimée.

- **Ecart-type** :

$$\hat{\sigma}^*(S) = \sqrt{\frac{n-1}{n} \sum_{i=1}^{n} (\hat{S}_{(-i)} - \hat{\mu}^*(S))^2}$$

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.

Onglet **Général**:

Données quantitatives : sélectionnez les échantillons de données pour lesquels vous voulez calculer les statistiques descriptives. Si des en-têtes ont été sélectionnés, veuillez vérifier que l'option « Libellés des échantillons » est activée.

Méthode : Choisissez la méthode de rééchantillonnage à utiliser.

- **Bootstrap** : activez cette option si vous désirez faire du bootstrap.
- **Aléatoire sans remise** : activez cette option si vous désirez effectuer un tirage aléatoire sans remise.
- **Jackknife** : activez cette option si vous désirez faire du jackknife.

Taille de l'échantillon : entrez la taille de l'échantillon dans le cas d'un tirage aléatoire avec remise.

Nombre d'échantillons : entrez le nombre d'échantillons dans le cas du bootstrap et d'un tirage aléatoire avec remise.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des échantillons : activez cette option si la première ligne des sélections (données quantitatives, qualitatives, sous-échantillons, poids) contient un libellé.

Poids : activez cette option si vous voulez pondérer les observations. Si vous n'activez pas cette option, les poids seront tous considérés comme valant 1. Les poids doivent être impérativement supérieurs ou égaux à 0. Si un en-tête de colonne a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

Onglet **Données manquantes** :

Supprimer les observations : activez cette option pour ne pas prendre en compte une observation dont l'une des données est manquante.

Estimer les données manquantes : activez cette option pour estimer les données manquantes en utilisant la moyenne de l'échantillon.

Onglet **Sorties**:

Données quantitatives : activez les options pour les statistiques descriptives que vous voulez calculer. Les différentes statistiques sont présentées dans la section description.

- **Toutes** : cliquez sur ce bouton pour tout sélectionner.
- **Aucune** : cliquez sur ce bouton pour tout désélectionner.

- **Affichage vertical** : activez cette option pour que le tableau des statistiques descriptives soit affiché verticalement (une ligne par statistique descriptive).

Intervalle de confiance : entrez la taille de l'intervalle de confiance désiré (en %)

Intervalle bootstrap standard : activez cette option pour afficher les bornes de l'intervalle de confiance bootstrap standard.

Intervalle des percentiles simples : activez cette option pour afficher les intervalles de confiance des percentiles simples.

Intervalle des percentiles corrigés par le biais : activez cette option pour afficher les Intervalles de confiance des percentiles corrigés par le biais.

Statistiques rééchantillonnées : activez cette option pour afficher l'ensemble des statistiques rééchantillonnées.

Données rééchantillonnées : activez cette option pour afficher l'ensemble des données rééchantillonnées.

Onglet **Graphiques** :

Histogrammes : activez cette option pour afficher les histogrammes des échantillons. Pour la distribution théorique, la fonction de densité est affichée.

- **Barres** : choisissez cette option pour afficher des histogrammes avec une barre pour chaque intervalle.
- **Lignes continues** : choisissez cette option pour afficher des histogrammes avec une ligne continue.

Histogrammes cumulés : activez cette option pour afficher les histogrammes cumulés des échantillons.

- **Basés sur l'histogramme** : choisissez cette option pour afficher des histogrammes cumulés basés sur la même définition d'intervalles que les histogrammes.
- **Fonction de répartition empirique** : choisissez cette option pour afficher des histogrammes cumulés qui correspondent en réalité à la fonction de répartition empirique de l'échantillon.

Ordonnées des histogrammes : choisissez quelle grandeur doit être utilisée pour les histogrammes : densité, effectif ou fréquence.

Résultats

Statistiques simples : dans ce tableau sont affichées pour tous les échantillons les statistiques descriptives suivantes : le nombre d'observations, le nombre de données manquantes, le nombre de données non manquantes, la moyenne, et l'écart-type (non biaisé).

Rééchantillonnage : dans ce tableau sont affichées pour chaque statistique descriptive sélectionnées : la moyenne, l'écart-type et les bornes des intervalles de confiance obtenus par rééchantillonnage.

Statistiques rééchantillonnées : dans ce tableau sont affichées les statistiques séléctionnées pour l'ensemble des B échantillons générés.

Données rééchantillonnées : dans ce tableau sont rassemblés les B échantillons générés à partir de l'échantillon original.

Histogrammes : les histogrammes sont affichés. Si vous le souhaitez, vous pouvez modifier la couleur des lignes, les échelles et les titres comme avec n'importe quel graphique Excel.

Statistiques descriptives pour les intervalles : dans ce tableau sont affichés pour chaque intervalle sa borne inférieure, sa borne supérieure, le nombre de valeurs de l'échantillon étant comprises dans l'intervalle (effectif), la fréquence (l'effectif divisé par l'effectif total de l'échantillon), et la densité (le rapport de la fréquence sur la taille de l'intervalle).

Exemple

Un exemple de rééchantillonnage est disponible sur le Centre d'aide XLSTAT à l'adresse

http://www.xlstat.com/demo-resamplef.htm

Bibliographie

Efron B. and Tibshirani R.J. (1993). An introduction to the bootstrap, Chapman & Hall / CRC.

Good P. (2006). Resampling methods. A guide to data analysis. Third Edition. Birkhäuser.

Matrices de similarité/dissimilarité (Corrélations, ...)

Utilisez cet outil pour calculer un indice de proximité entre les lignes ou les colonnes d'un tableau de données. Le cas le plus classique d'utilisation de cet outil est le calcul d'une matrice de corrélation ou de covariance entre des variables quantitatives.

Dans cette section :

Description

Cet outil propose un nombre important de mesures de proximité entre une série d'objets, qu'il s'agisse de lignes (en principe des observations) ou de colonnes (en principe des variables).

Le coefficient de corrélation est une mesure de similarité des variables : plus des variables sont similaires, plus le coefficient de corrélation est élevé.

Similarités et dissimilarités

La mesure de la proximité entre deux objets peut se faire en mesurant à quel point ils sont semblables (similarité) ou dissemblables (dissimilarité).

Les indices proposés dépendent de la nature de données :

- Données quantitatives :

Les indices de **similarité** proposés pour des calculs à partir de données quantitatives sont les suivants : Cosinus, Covariance (n-1), Covariance (n), Indice de Gower, Inertie, Coefficient de corrélation de Kendall, Coefficient de corrélation partielle, Coefficient de corrélation de Pearson, Similarité générale, Coefficient de corrélation de Spearman.

Les indices de **dissimilarité** proposés pour des calculs à partir de données quantitatives sont les suivants : Distance de Bhattacharya, Distance de Bray et Curtis, Distance de Canberra, Distance de Chebychev, Distance du Khi², Métrique du Khi², Distance de la corde, Distance de la corde au carré, Distance euclidienne, Distance euclidienne carré, Distance géodésique, Dissimilarité de Kendall, Distance de Mahalanobis, Distance de Manhattan, Dissimilarité de Pearson, Dissimilarité générale, Dissimilarité de Spearman.

- Données binaires :

Les indices de **similarité** et de **dissimilarité** (par simple transformation) proposés pour des calculs à partir de données binaires sont les suivants : Cooccurrence, Indice de Dice (aussi appelé indice de Sorensen), Indice de Jaccard (1), Indice de Jaccard (2), Indice de Rand, Indice de Rand ajusté, Indice de Kulczinski, Phi de Pearson, Similarité générale, Indice d'Ochiai, Indice de Rogers & Tanimoto, Indice de Sokal & Michener (*simple matching coefficient*), Indice de Sokal & Sneath(1), Indice de Sokal & Sneath(2).

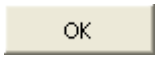
- Données qualitatives :

Les indices de **similarité** proposés pour des calculs à partir de données qualitatives sont les suivants : Cooccurrence, Similarité générale.

L'indice de **dissimilarité** proposé pour des calculs à partir de données qualitatives est le suivant : Dissimilarité générale.

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.

Onglet **Général**:

Données : sélectionnez un tableau comprenant N objets décrits par P descripteurs. Si des en-têtes de colonnes ont été sélectionnés, veuillez vérifier que l'option « Libellés des colonnes » est activée.

Type de données : choisissez le type des données sélectionnées.

Remarque : dans le cas où le type de données choisi est « Qualitatives », quelque soit leur type réel, les données sont considérées comme qualitatives.

Poids des lignes : activez cette option si vous voulez pondérer les lignes. Si vous n'activez pas cette option, les poids seront tous considérés comme valant 1. Les poids doivent être impérativement supérieurs ou égaux à 0. Si un en-tête de colonne a été sélectionné, veuillez vérifier que l'option « Libellés des colonnes » est activée.

Type de proximité : **similarités / dissimilarités** : choisissez le type de proximité à utiliser. Le type de données et le type de proximité déterminent la liste des indices possibles pour le calcul de la matrice de proximité.

Remarque : pour calculer un coefficient de corrélation classique (aussi appelé coefficient de corrélation de Pearson), vous devez sélectionner le type de données « quantitatives », « similarités », et le « Coefficient de corrélation de Pearson ».

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des colonnes : activez cette option si la première ligne des données sélectionnées (Tableau observations/variables, libellés des lignes, poids des lignes, poids des colonnes) contient un libellé.

Libellés des lignes : activez cette option si vous voulez utiliser des libellés d'observations pour l'affichage des résultats. Si l'option « Libellés des colonnes » est activée, la première cellule de la sélection doit comprendre un en-tête. Si vous n'activez pas cette option, des libellés seront automatiquement créés (Obs1, Obs2, …).

Calculer les proximités pour les :

Colonnes : activez cette option si vous voulez mesurer la proximité entre les colonnes.

Lignes : activez cette option si vous voulez mesurer la proximité entre les lignes.

Onglet **Données manquantes** :

Ne pas accepter les valeurs manquantes : activez cette option pour que XLSTAT empêche la poursuite des calculs si des valeurs manquantes sont détectées.

Supprimer les observations : activez cette option pour supprimer les observations comportant des données manquantes.

Suppression par paires : activez cette option pour supprimer les observations comportant des données manquantes uniquement lorsque les variables impliquées dans les calculs comportent des données manquantes. Par exemple lors du calcul d'une corrélation entre deux variables, une observation ne sera ignorée que si la donnée correspondant à l'une des deux variables est manquante.

Estimer les données manquantes : activez cette option pour estimer les données manquantes avant le début des calculs.

- **Moyenne ou mode** : activez cette option pour estimer les données manquantes en utilisant la moyenne (variables quantitatives) ou le mode (variables qualitatives) pour les variables correspondantes.
- **Plus proche voisin** : activez cette option pour estimer les données manquantes d'une observation en recherchant le plus proche voisin de l'observation.

Onglet **Sorties**:

Statistiques descriptives : activez cette option pour afficher les statistiques descriptives pour les variables sélectionnées.

Identifier les objets similaires : activez cette option pour identifier dans la matrice de proximité les objets similaires.

Lister les objets similaires : activez cette option pour afficher la liste des objets similaires.

Seuil de dissimilarité : entrez la valeur seuil de l'indice à partir de laquelle vous considérez que les objets sont similaires. Si l'indice choisi est une similarité, les données seront considérées comme étant similaires si elles sont supérieures à cette valeur. Si vous avez choisi un indice de dissimilarité, les données seront considérées comme étant similaires si elles sont inférieures à cette valeur.

Alpha de Cronbach : activez cette option pour calculer l'alpha de Cronbach.

Test de sphéricité de Bartlett : activez cette option pour calculer le test de sphéricité de Bartlett (uniquement dans le cas de la corrélation de Pearson ou de la covariance).

Niveau de signification (%) : entrez le niveau de signification pour le test de sphéricité.

Résultats

Statistiques simples : dans ce tableau sont affichées les statistiques descriptives des échantillons.

Matrice de proximité : dans ce tableau sont affichées les proximités entre les objets pour l'indice choisi. Si l'option « Identifier les objets similaires a été activée » et que le seuil de dissimilarité est dépassé, les valeurs correspondant à des objets similaires sont affichées en gras.

Liste des objets similaires : si l'option « lister les objets similaires » est activée et qu'au moins une paire d'objets a une dissimilarité au-delà de ce seuil, la liste des objets similaires est affichée.

Exemple

Un exemple montrant comment calculer une matrice de dissimilarité est disponible sur le Centre d'aide XLSTAT :

http://www.xlstat.com/demo-mdsf.htm

Bibliographie

Everitt B.S., Landau S. and Leese M. (2001). Cluster Analysis (4th edition). Arnold, London.

Gower J.C. and P. Legendre (1986). Metric and Euclidean properties of dissimilarity coefficients. *Journal of Classification*, **3**, 5-48.

Jobson J.D. (1992). Applied Multivariate Data Analysis. Volume II: Categorical and Multivariate Methods. Springer-Verlag, New York.

Legendre P. and Legendre L. (1998). Numerical Ecology. Second English Edition. Elsevier, Amsterdam.

Sokal R.R. and Rohlf F.J. (1995). Biometry. The Principles and Practice of Statistics in Biological Research. Third edition. Freeman, New York.

Corrélation bisérielle

Utilisez cet outil pour calculer la corrélation bisérielle entre d'une part une ou plusieurs variables quantitatives et d'autre part une ou plusieurs variables qualitatives binaires.

Dans cette section :

Description

Cet outil permet de calculer la corrélation bisérielle entre d'une part une ou plusieurs variables quantitatives et d'autre part une ou plusieurs variables qualitatives binaires. La corrélation bisérielle, introduite par Pearson en 1909, entre une variable quantitative et une variable binaire est donnée par :

$$r = \frac{(\hat{\mu}_2 - \hat{\mu}_1)}{\hat{\sigma}_n}\sqrt{p_1 p_2}$$

Où $\hat{\mu}_1$ et $\hat{\mu}_2$ sont les moyennes estimées pour la variable quantitative et correspondant respectivement à la première et à la seconde modalité de la variable qualitative binaire, $\hat{\sigma}_n$ est l'écart-type biaisé (division par n) estimé sur l'ensemble des données (toutes modalités confondues), p_1 et p_2 sont les proportions associées aux deux modalités de la variable qualitative ($p_1 + p_2 = 1$). Comme pour le coefficient de corrélation de Pearson, le coefficient r est compris entre -1 et 1. La valeur 0 correspond au cas où il n'y a pas d'association, les moyennes de la variable quantitative pour les 2 modalités de la variable qualitative étant égales.

XLSTAT permet de tester si le r obtenu est significativement différent de 0 ou non.

Pour le test bilatéral, les hypothèses nulle H0 et alternative Ha sont les suivantes :

- H0 : $r = 0$
- Ha : $r \neq 0$

Pour le test unilatéral à gauche, les hypothèses sont les suivantes :

- H0 : $r = 0$
- Ha : $r < 0$

Pour le test unilatéral à droite, les hypothèses sont les suivantes :

- H0 : $r = 0$
- Ha : $r > 0$

Deux méthodes de calcul sont proposées par XLSTAT pour le calcul de la p-value. L'utilisateur peut choisir entre un calcul basé sur une approximation et un calcul basé sur des rééchantillonnages Monte Carlo. La seconde méthode est recommandée car plus fiable et rapide.

Pour le calcul de la p-value par approximation, on utilise le résultat suivant :

Si n est la taille de l'échantillon complet, la statistique définie par

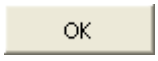
t = $r\sqrt{\frac{n-2}{1-r^2}}$

suit une loi de Student à $n - 2$ degrés de liberté sous l'hypothèse nulle.

Remarque : la fonction XLSTAT_Biserial permet de calculer la corrélation bisérielle directement dans une feuille de calcul.

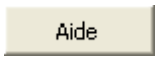
Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.



: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.

Onglet **Général**:

Variables quantitatives : activez cette option pour sélectionner les variables quantitatives. Si des en-têtes ont été sélectionnés, veuillez vérifier que l'option « Libellés des variables » est activée.

Variables qualitatives : activez cette option pour sélectionner les variables qualitatives. Si des en-têtes ont été sélectionnés, veuillez vérifier que l'option « Libellés des variables » est activée.

Poids : activez cette option si vous voulez pondérer les observations. Si vous n'activez pas cette option, les poids seront tous considérés comme valant 1. Les poids doivent être impérativement supérieurs ou égaux à 0. Si un en- tête de colonne a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

Modalité témoin : choisissez quelle modalité est la modalité témoin, c'est à dire quel est le groupe 2 dans les calculs ci-dessus.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des variables : activez cette option si la première ligne des données sélectionnées contient un libellé.

Onglet **Options** :

Hypothèse alternative : choisissez l'hypothèse alternative à utiliser pour le test (voir la section description).

Niveau de signification (%) : entrez le niveau de signification à utiliser pour les différents tests (valeur par défaut : 5%).

p-value asymptotique : activez cette option pour calculer la p-value basée sur la distribution asymptotique de la statistique t (voir la section description).

Méthode Monte Carlo : activez cette option pour calculer la p-value en utilisant des simulations Monte Carlo. Entrez alors le nombre de simulations et le temps maximum à consacrer aux calculs.

Onglet **Données manquantes** :

Ne pas accepter les valeurs manquantes : activez cette option pour que XLSTAT empêche la poursuite des calculs si des valeurs manquantes sont détectées.

Supprimer les observations : activez cette option pour supprimer les observations comportant des données manquantes.

- **Pour la variable correspondante** : activez cette option pour ne pas prendre en compte une observation dont l'une des données est manquante, uniquement pour les variables pour lesquels une donnée est manquante.
- **Pour toutes les variables** : activez cette option pour ne pas prendre en compte une observation dont l'une des données est manquante, pour toutes les variables sélectionnées.

Onglet **Sorties**:

Statistiques descriptives : activez cette option pour afficher les statistiques descriptives pour les variables sélectionnées.

Résultats

Statistiques simples : dans ce tableau sont affichées les statistiques descriptives des échantillons.

La valeur de la corrélation est ensuite donnée pour chaque couple (variable quantitative,variable qualitative), éventuellement avec le calcul de la p-value associée si une p-value est demandée. Les détails ne sont donnés que lorsque la corrélation bisérielle est calculée sur 1 variable quantitative et une variable qualitative.

Exemple

Un exemple de calcul de corrélation bisérielle est disponible sur le Centre d'aide XLSTAT :

http://www.xlstat.com/demo-biserialf.htm

Bibliographie

Chmura Kraemer H. (1982). Biserial Correlation, Encyclopaedia of Statistical Sciences, Volume 1, Wiley, 276-279.

Pearson K. (1909). On a New Method of Determining Correlation between a measured Character A and a Character B, of which only the Percentage of cases wherein B exceeds (or falls short of) a given Intensity is recorded for each grade of A. *Biometrika*, **7**, 96-105.

Richardson M.W. and Stalnaker J.M. (1933). A note on the use of bi-serial r in test research. *Journal of General Psychology,* **8**, 463-465.

Statistiques de multicolinéarité

Utilisez cet outil pour identifier des multicolinéarités entre vos variables.

Dans cette section :

Description

On dit que des variables sont multicolinéaires s'il existe une relation linéaire entre elles. C'est une extension du cas simple de la colinéarité entre deux variables. Par exemple, pour trois variables X_1, X_2, X_3, on dira qu'elles sont multicolinéaires si on peut écrire :

$$X_1 = aX_2 + bX_3$$

où a et b sont deux nombres réels.

Si l'Analyse en Composantes Principales (ACP) permet de détecter la présence de multicolinéarités au sein des données (un nombre de facteurs non nuls inférieur au nombre de variables indique la présence d'une multicolinéarité), elle ne permet pas d'identifier les variables qui en sont responsables.

Pour détecter les multicolinéarités et identifier les variables impliquées dans des multicolinéarités, on effectue des régressions linéaires de chacune des variables en fonction des autres. On calcule ensuite :

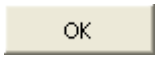
- le R² de chacun des modèles. Si le R² vaut 1, alors il existe une relation linéaire entre la variable dépendante du modèle (le Y) et les variables explicatives (les X).
- la **tolérance** pour chacun des modèles. La tolérance vaut (1-R²). Elle est utilisée dans plusieurs méthodes (régression linéaire, régression logistique, analyse factorielle discriminante) comme un critère de filtrage des variables. Si une variable a une tolérance inférieure à un seuil fixé (la tolérance est calculée en prenant en compte les variables déjà utilisées dans le modèle), on ne la laisse pas entrer dans le modèle car sa contribution est négligeable et elle risquerait d'entraîner les problèmes numériques.
- le VIF (Variance Inflation Factor) qui est égal à l'inverse de la tolérance.

Il peut être utile de détecter des multicolinéarités au sein d'un groupe de variables notamment dans les cas suivants :

- pour identifier des structures dans les données et en tirer des décisions opérationnelles (par exemple, arrêter de mesurer une variable sur une chaîne de fabrication car elle est fortement liée à d'autres qui sont aussi mesurées) ;
- pour éviter des problèmes numériques lors de certains calculs. Certaines méthodes utilisent des inversions de matrices. L'inverse d'une matrice $(p \times p)$ ne peut être calculé que si elle est de rang p (ou régulière). Si elle est de rang inférieur, autrement dit s'il existe des relations linéaires entre ses colonnes, alors elle est singulière et non inversible.

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.

Onglet **Général**:

Tableau observations/variables : sélectionnez un tableau comprenant N objets décrits par P variables. Si des en-têtes de colonnes ont été sélectionnés, veuillez vérifier que l'option « Libellés des variables » est activée.

Libellés des variables : activez cette option si la première ligne des données sélectionnées contient un libellé.

Poids : activez cette option si vous voulez pondérer les observations. Si vous n'activez pas cette option, les poids seront tous considérés comme valant 1. Les poids doivent être

impérativement supérieurs ou égaux à 0. Si un en- tête de colonne a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Onglet **Données manquantes** :

Ne pas accepter les valeurs manquantes : activez cette option pour que XLSTAT empêche la poursuite des calculs si des valeurs manquantes sont détectées.

Supprimer les observations : activez cette option pour supprimer les observations comportant des données manquantes.

Onglet **Sorties**:

Statistiques descriptives : activez cette option pour afficher les statistiques descriptives pour les variables sélectionnées.

Corrélations : activez cette option pour afficher la matrice de corrélation.

R^2 : activez cette option pour afficher les R^2.

Tolérance : activez cette option pour afficher les tolérances.

VIF : activez cette option pour afficher les VIF.

Onglet **Graphiques** :

Diagrammes en bâtons : activez cette option pour afficher les diagrammes en bâton des statistiques suivantes :

- R^2
- Tolérance
- VIF

Résultats

Les résultats comprennent les statistiques descriptives des variables sélectionnées, la matrice de corrélation des variables et les statistiques de multicolinéarité (R^2, Tolérance et VIF). Des diagrammes en bâtons permettent de repérer les variables les plus multi-corrélées à d'autres.

Lorsque la tolérance vaut 0, le VIF a une valeur infinie et n'est pas affiché.

Exemple

Bibliographie

Belsley D.A., Kuh E. and Welsch R.E. (1980). Regression Diagnostics, Identifying Influential Data and Sources of Collinearity. Wiley, New York.

Analyse de la fiabilité

L'analyse de fiabilité permet de caractériser des échelles de mesure composées de divers éléments (par exemple des questions dans le cas d'un questionnaire). La procédure utilisée calcule plusieurs mesures qui permettent d'évaluer la fiabilité de l'échelle et propose également des informations sur les relations entre les différents éléments.

Dans cette section :

Description

L'analyse de fiabilité est utilisée dans différents domaines et plus particulièrement dans les sciences sociales. La terminologie est issue de la psychométrie. On définit ainsi un test, qui est lui-même composé d'un ensemble de questions. Les questions sont appelées éléments. Ces éléments sont regroupés en construits homogènes (construct en anglais) aussi appelés facteur, échelle de mesure, variable latente ou concept. Par exemple, l'aptitude graphique est un facteur dont on voudra mesurer le niveau au travers d'une échelle de mesure. Le but de l'analyse de fiabilité est de vérifier que l'échelle de mesure est fiable, autrement dit que les différentes questions d'un construit sont cohérentes, qu'elles mesurent bien la même chose. Par exemple, si l'on s'intéresse à l'aptitude graphique, une question de calcul mental nuirait à la cohérence de l'échelle.

Pour le statisticien les questions sont des variables souvent mesurées sur des échelles de type Likert (réponses graduées). Les résultats d'un test effectué auprès d'un ensemble d'individus sont consignés dans un tableau individus/variables. Ces méthodes pouvant être utilisées dans d'autres domaines comme le contrôle qualité, XLSTAT désigne ces tableaux comme des tableaux observations/variables.

Les méthodes implémentées dans XLSTAT permettent d'estimer la cohérence interne d'une échelle de mesure, autrement dit de voir si les résultats des différentes questions, censées mesurer le même phénomène, sont cohérents mais aussi la corrélation entre deux tests administrés aux mêmes individus à deux moments différents.

L'analyse, dite « interne », permet d'une part de déterminer quels éléments d'un questionnaire sont corrélés en fournissant un indice général de la cohérence interne de l'échelle globale et, d'autre part, d'identifier les éléments inutiles et donc de les exclure de l'échelle.

L'analyse, dite en « deux parties », mesure quant à elle l'équivalence en deux parties d'un test (fiabilité des formes parallèles). Dans ce cas, on administre deux piles d'éléments mesurant la même chose et ce avec le même instrument et avec les mêmes personnes.

L'indice alpha de Cronbach est une mesure de la cohérence interne d'un test ou, autrement dit, une mesure de la fiabilité de l'échelle.

Cet indice (représenté par la lettre grecque « α ») est l'équivalent mathématique de l'estimation de la moyenne de toutes les corrélations entre deux parties égales de l'échelle.

L'alpha de Cronbach est donné par :

$$\alpha = \frac{\sum_{h \neq h'} \mathrm{cov}(x_h, x_{h'})}{\mathrm{var}\left(\sum_h x_h\right)} \times \frac{p}{p-1}$$

Avec x_h l'élément h de l'échelle et p le nombre total d'éléments.

Il est recommandé d'avoir un coefficient alpha minimum compris entre 0,65 et 0,8 (ou plus); ceux inférieurs à 0,5 sont habituellement inacceptables.

XLSTAT fournit également le calcul de l'alpha de Cronbach dit « standardisé ». Celui-ci équivaut à la fiabilité qui serait obtenue si l'ensemble des réponses à chaque question avait été standardisé avant calcul.

L'alpha de Cronbach standardisé est donné par :

$$\alpha_{std} = \frac{\sum_{h \neq h'} \mathrm{cor}(x_h, x_{h'})}{p + \sum_{h \neq h'} \mathrm{cor}(x_h, x_{h'})} \times \frac{p}{p-1}$$

Indices de Guttman

Les bornes inférieures de Guttman (lambda 1-6) sont un ensemble de six coefficients, L1 à L6 servant eux aussi à estimer la fiabilité interne (L1, L3, L5, L6) ou en deux parties d'un test (L2, L4) :

L1: Un coefficient intermédiaire utilisé dans le calcul des autres lambdas.

L2: Estimation de la corrélation inter-score dans le cas de mesures parallèles. Il est plus complexe que l'alpha de Cronbach et représente mieux la vraie fiabilité du test.

L3: Equivalent à l'alpha de Cronbach.

L4: Fiabilité en deux parties de Guttman (Cf. description ci-dessous)

L5: Recommandé lorsqu'un seul item covarie fortement avec les autres, lesquels ne présentent pas de covariances élevées les uns avec les autres.

L6: Recommandé lorsque les corrélations inter-éléments sont faibles par rapport aux coefficients de déterminations item vs items restant (devient un meilleur estimateur lorsque le nombre d'items devient important).

Fiabilité en deux parties de Spearman-Brown (modèle en deux parties) :

Une autre manière de calculer la fiabilité d'une échelle de mesure consiste à la fractionner aléatoirement en deux parties. Si l'échelle est parfaitement fiable, nous devons nous attendre à ce que les deux moitiés soient parfaitement corrélées (c'est-à-dire $R = 1$). Une fiabilité imparfaite conduit à des corrélations imparfaites. Nous pouvons estimer cette fiabilité de l'échelle grâce au coefficient par moitié de Spearman-Brown :

$$Y = \frac{2R}{1 + R}$$

Dans cette formule, Y est le coefficient de fiabilité par moitié et R représente la corrélation entre les deux parties de l'échelle.

Lorsque les deux parties ont des tailles différentes, une estimation plus précise de la fiabilité est utilisée (Formule de Horst), laquelle est définie comme suit :

$$H = \frac{-R^2 + \sqrt{R^4 + 4R^2(1 - R^2)k_1k_2/k}}{2(1 - R^2)k_1k_2/k}$$

Dans cette formule, H est le coefficient de fiabilité par moitié de Horst, R représente la corrélation entre les deux moitiés de l'échelle, k_1 le nombre d'items de la première partition, k_2 le nombre d'items de la seconde partition et k le nombre total d'items de l'échelle.

Fiabilité en deux parties de Guttman (modèle en deux parties) :

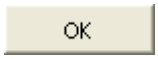
Le coefficient de fiabilité par partie de Guttman est semblable au coefficient de fiabilité par partie de Spearman-Brown, mais il ne considère pas que les fiabilités ou bien les variances sont égales dans les deux parties (Tau- equivalence). Il est calculé comme suit :

$$G = \frac{2 \cdot (S_p^2 - S_{p_1}^2 - S_{p_2}^2)}{S_p^2}$$

Dans cette formule, G est le coefficient de fiabilité par moitié de Guttman (L4) etS_p,S_{p_1},S_{p_2} sont les variances respectives de l'échelle totale et des sous-parties du test.

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.

Onglet **Général** :

Tableau observations/éléments : sélectionnez un tableau comprenant les observations. Si des en-têtes de colonne ont été sélectionnées pour les éléments, veuillez vérifier que l'option « Libellés des variables » est activée.

Tableau observations/éléments (1) : sélectionnez un tableau comprenant les observations de la première moitié du test (modèle en deux parties). Si des en-têtes de colonne ont été sélectionnées pour les éléments, veuillez vérifier que l'option « Libellés des variables » est activée.

Tableau observations/éléments (2) : sélectionnez un tableau comprenant les observations de la seconde moitié du test (modèle en deux parties). Si des en- têtes de colonne ont été sélectionnées pour les éléments, veuillez vérifier que l'option « Libellés des variables » est activée.

Poids : activez cette option si vous voulez pondérer les observations. Si vous n'activez pas cette option, les poids seront tous considérés comme valant 1. Les poids doivent être impérativement supérieurs ou égaux à 0. Si un en- tête de colonne a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

Type de fiabilité : choisissez le type de fiabilité à utiliser pour les calculs (voir la section description pour plus de détails).

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des variables : activez cette option si la première ligne (ou colonne en mode lignes) des données sélectionnées (tableau observations/variables et poids) contient un libellé.

Onglet **Options** :

Enumération : activez cette option pour rechercher la partition optimale (modèle en deux parties) maximisant l'indice de Guttman L4 de l'échelle donnée en entrée du test. Cette recherche s'effectue en testant l'ensemble des combinaisons en deux parties du test initial.

- **Temps maximum (s)** : activez cette option pour fixer un délai maximum en secondes pour la recherche de l'indice de Guttman L4 optimal (valeur max).
- **Rapide** : Emploi un algorithme de recherche du maximum de Guttman L4 fournissant une solution optimale en un temps réduit (préférable lorsque le nombre d'items du test devient important)

Onglet **Données manquantes** :

Ne pas accepter les valeurs manquantes : activez cette option pour que XLSTAT empêche la poursuite des calculs si des valeurs manquantes sont détectées.

Supprimer les observations : activez cette option pour supprimer les observations comportant des données manquantes.

Suppression par paires : activez cette option pour supprimer les observations comportant des données manquantes uniquement lorsque les variables impliquées dans les calculs comportent des données manquantes. Par exemple, lors du calcul d'une corrélation entre deux variables, une observation ne sera ignorée que si la donnée correspondant à l'une des deux variables est manquante.

Estimer les données manquantes : activez cette option pour estimer les données manquantes avant le début des calculs.

- **Moyenne ou mode** : activez cette option pour estimer les données manquantes en utilisant la moyenne (variables quantitatives) ou le mode (variables qualitatives) pour les variables correspondantes.
- **Plus proche voisin** : activez cette option pour estimer les données manquantes d'une observation en recherchant le plus proche voisin de l'observation.

Onglet **Sorties** :

Statistiques descriptives : activez cette option pour afficher les statistiques descriptives pour les éléments sélectionnés ainsi que pour les échelles.

Statistiques des éléments supprimés : activer cette option pour calculer et afficher les statistiques sur la comparaison de chaque élément à l'échelle composée des autres éléments. Ces statistiques incluent la moyenne et la variance de l'échelle si l'élément a été supprimé, la

corrélation entre l'élément et l'échelle composée des autres éléments, l'alpha de Cronbach si l'élément a été supprimé de l'échelle, l'indice de Guttman L6 si l'élément a été supprimé de l'échelle et le coefficient de détermination (R^2) entre l'élément supprimé et l'échelle composée des autres éléments.

Matrice de corrélation : activez cette option pour afficher la matrice de corrélation correspondant à la corrélation de Pearson appliquée à la table d'observations en entrée.

Matrice de covariance : activez cette option pour afficher la matrice de covariance appliquée à la table d'observations en entrée.

Statistiques alpha de Cronbach : activez cette option pour afficher l'alpha de Cronbach brute et standardisé de l'échelle globale et de chaque sous-partie lorsque l'option « modèle en deux parties » est sélectionnée.

Statistiques de Guttman : activez cette option pour afficher les différents indices de Guttman

- **Afficher la meilleure fraction** : activez cette option pour afficher chacune des partitions correspondant à un coefficient de Guttman L4 optimal.

Statistiques de modèle en deux parties : activez cette option pour afficher les différents indices correspondant à un modèle de fiabilité en deux parties (corrélation entre les deux moitiés de l'échelle, coefficient de Spearman-Brown ou coefficient de Horst, coefficient de fiabilité par moitié de Guttman (L4) pour la partition donnée en entrée)

Onglet **Graphiques** :

Cartes des corrélations : plusieurs représentations d'une matrice des corrélations vous sont proposées.

- L'option « **Echelle bleu-rouge** » vous permet de représenter les corrélations faibles par des couleurs froides (bleu pour les corrélations proches de -1) et les corrélations élevées par des couleurs chaudes (rouge pour les corrélations proches de 1).

- L'option « **Noir et blanc** » vous permet soit de représenter en noir les corrélations positives et en blanc les corrélations négatives (la diagonale de 1 est représentée en gris), soit de représenter en noir les corrélations significativement non nulles, et en blanc les corrélations non significativement différentes de 0.

- L'option « **Motifs** » vous permet de représenter les corrélations positives par des traits montant de gauche à droite, et les corrélations négatives par des traits montant de droite à gauche. Plus la corrélation est élevée en valeur absolue, plus les traits sont espacés.

Résultats

La matrice de corrélation et les statistiques descriptives échelle(s)/éléments sont affichées.

Les statistiques de l'alpha de Cronbach/Guttman permettent de connaitre la fiabilité de l'échelle saisie en entrée, tandis que les statistiques descriptives des éléments supprimés renseignent sur l'influence du retrait de chaque élément sur la fiabilité globale de l'échelle.

L'analyse de la fiabilité interne permet un calcul de la meilleure partition, utilisable dans le cas où une analyse en deux parties est envisagée par la suite.

La carte des corrélations permet quant à elle de mettre en évidence d'éventuelles structures dans les corrélations, ou d'identifier rapidement les éléments ayant des corrélations intéressantes.

Exemple

Un exemple portant sur des données provenant du site web Personality Tests est disponible en permanence sur le Centre d'aide XLSTAT :

http://www.xlstat.com/demo-reliabilityf.htm

Bibliographie

Cronbach L. J. (1951). Coefficient Alpha and the internal structure of test. *Psychometrika*, **16** (3), 297-334.

Guttman L (1945) A basis for analyzing test–retest reliability. Psychometrika 10:255–282

Tableau de contingence (statistiques descriptives)

Utilisez cet outil pour calculer des statistiques descriptives sur un tableau de contingence. Un test d'indépendance du khi² entre les lignes et les colonnes peut être calculé.

Dans cette section :

Description

Un tableau de contingence est une manière efficace de résumer la relation entre deux variables qualitatives V_1 et V_2. Un tableau de contingence a la structure suivante :

$V_1 \setminus V_2$	**Modalité 1**	**…**	**Modalité j**	**…**	**Modalité m_2**
Modalité 1	$n(1,1)$		$n(1,j)$	…	$n(1,m_2)$
…	…	…	…	…	…
Modalité i	$n(i,1)$	…	$n(i,j)$	…	$n(i,m_2)$
…	…	…	…	…	…
Modalité m_1	$n(m_1,1)$	…	$n(m_1,j)$	…	$n(m_1,m_2)$

où $n(i,j) = n_{ij}$ est la fréquence des observations présentant à la fois la caractéristique i pour la variable V_1, et la caractéristique j pour la variable V_2.

La distance du khi² a été proposée pour mesurer la distance entre les modalités. La somme de ces distances pour l'ensemble des cases du tableau donne la statistique du khi² qui suit asymptotiquement une loi du khi² à $(m_1 - 1)(m_2 - 1)$ degrés de liberté. Cette statistique permet de tester l'hypothèse d'indépendance entre les lignes et les colonnes du tableau de contingence.

La notion d'inertie inspirée de la physique est utilisée en Analyse Factorielle des Correspondances. L'inertie d'un nuage de points est la moyenne pondérée des carrés des distances au centre de gravité. L'inertie totale du nuage des modalités est donnée par :

$$\phi^2 = \frac{\chi^2}{n} = \sum_{i=1}^{m_1}\sum_{j=1}^{m_2} \frac{\left(\frac{n_{ij}}{n} - \frac{n_{i.}n_{.j}}{n^2}\right)^2}{\frac{n_{i.}n_{.j}}{n^2}}, \text{ avec } n_{i.} = \sum_{j=1}^{m_2} n_{ij} \text{ et } n_{.j} = \sum_{j=1}^{m_1} n_{ij}$$

où n est la somme des fréquences du tableau de contingence. On voit ici que l'inertie totale est proportionnelle à la statistique du khi² de Pearson mesurée sur le tableau de contingence.

Intervalle de confiance bootstrap

XLSTAT vous permet de calculer des intervalles de confiance bootstrap autour des valeurs théoriques obtenus sur les tableaux de contingence. Ceux-ci permettent une alternative non paramétrique au test du Khi² par case.

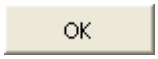
La procédure est la suivante :

1. Construction d'un jeu de données composé de deux variables qualitatives à partir du tableau.
2. Tirage aléatoire avec remise (bootstrap) de N observations indépendamment dans chaque variable.
3. Construction d'un tableau de contingence à partir du nouveau jeu de données et obtention des fréquences théoriques pour chaque case du tableau de contingence.
4. Les étapes 2 et 3 sont répétées autant de fois que demandé par l'utilisateur.
5. Les moyennes, écarts-types, intervalles de confiance et intervalles de confiance percentiles sont obtenus.

Les paires dont la valeur observée est hors de l'intervalle de confiance traduisent une relation entre les deux modalités (Amiri *et al.*, 2011).

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les

variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.

Onglet **Général**:

Tableau de contingence : sélectionnez un tableau croisé, avec les fréquences correspondant aux différentes catégories de deux variables qualitatives. Si les libellés des lignes et des colonnes du tableau ont été sélectionnés, veillez à ce que l'option « libellés inclus » soit activée.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés inclus : activez cette option si la première ligne et la première colonne des données sélectionnées contient un libellé.

Onglet **Options**:

Test du khi² : activez cette option pour effectuer le test du khi².

Niveau de signification (%) : entrez le niveau de signification à utiliser pour les différents tests (valeur par défaut : 5%).

Onglet **Données manquantes** :

Ne pas accepter les valeurs manquantes : activez cette option pour que XLSTAT empêche la poursuite des calculs si des valeurs manquantes sont détectées.

Remplacer les valeurs manquantes par 0 : activez cette option si vous considérez que les valeurs manquantes sont équivalentes à des 0.

Remplacer les valeurs manquantes par l'espérance : activez cette option si vous souhaitez remplacer les valeurs manquantes par leur espérance. L'espérance d'une valeur manquante est donnée par :

$$E(n_{ij}) = \frac{n_{i.}n_{.j}}{n}$$

où $n_{i.}$ est la somme sur les colonnes pour la ligne i, $n_{.j}$ est la somme sur les lignes pour colonne j, et n est l'effectif total avant remplacement des valeurs manquantes.

Onglet **Sorties**:

Liste des combinaisons : activez cette option pour afficher la liste des différentes combinaisons possibles des deux variables qualitatives, ainsi que les effectifs correspondants.

Tableau de contingence : activez cette option pour afficher le tableau de contingence.

Inertie par case : activez cette option pour afficher les inerties correspondant à chacune des cellules du tableau de contingence.

Khi² par case : activez cette option pour afficher les Khi² correspondant à chacune des cellules du tableau de contingence.

Significativité par case : activez cette option pour afficher un tableau indiquant, pour chaque case, si la valeur observée est égale (=), inférieure (<) ou supérieure (>) à la valeur théorique, et pour effectuer un test (test exact de Fisher sur un tableau 2×2 ayant le même effectif total que le tableau complet, et les mêmes sommes marginales pour la case en question), afin de déterminer si l'écart à la valeur théorique est significatif ou non.

Effectifs observés : activez cette option pour afficher le tableau des effectifs observés. Ce tableau est presque identique au tableau de contingence, la différence venant des sommes marginales pour les lignes et les colonnes.

Effectifs théoriques : activez cette option pour afficher le tableau des effectifs théoriques estimés à partir des sommes marginales.

Proportions ou pourcentages / Ligne : activez cette option pour afficher le tableau des proportions ou pourcentages par ligne qui correspondent aux effectifs observés divisés par les sommes marginales des lignes.

Proportions ou pourcentages / Colonne : activez cette option pour afficher le tableau des proportions ou pourcentages par colonne qui correspondent aux effectifs observés divisés par les sommes marginales des colonnes.

Proportions ou pourcentages / Total : activez cette option pour afficher le tableau des proportions ou pourcentages calculés comme les effectifs observés divisés par l'effectif total.

Données brutes : activez cette option pour afficher le tableau des données brutes, c'est à dire le tableau observations/variables table, ayant n lignes et 2 colonnes.

Onglet **Graphiques** :

Vue 3D du tableau de contingence / du tableau croisé : activez cette option pour afficher le diagramme en bâton en 3 dimensions correspondant au tableau de contingence ou au tableau croisé.

Tableau de contingence : activez cette option pour afficher le graphique associé au tableau de contingence.

Proportions ou pourcentages / Ligne : activez cette option pour afficher le graphique associé au tableau des proportions ou pourcentages par ligne.

Proportions ou pourcentages / Colonne : activez cette option pour afficher le graphique associé au tableau des proportions ou pourcentages par colonne.

Options des graphiques :

- **Type de graphique** :
    - **Groupé** : choisissez cette option pour afficher les graphiques sous forme de barres regroupées par modalité.
    - **Barres empilées** : choisissez cette option pour afficher les graphiques sous forme de barres empilées. Cela permet de comparer les effectifs ou les fréquences des sous-échantillons à ceux d'un échantillon complet.
- **Diagrammes en bâtons** :
    - **Effectifs** : choisissez cette option pour afficher l'effectif correspondant à chaque barre.
    - **Pourcentages** : choisissez cette option pour afficher le % de population correspondant à chaque barre

Bibliographie

Amiri S. and von Rosen D. (2011). On the efficiency of bootstrap method into the analysis contingency table. Computer methods and programs in biomedicine, **104(2)**, 182-187.

Générateur de tableaux croisés

Utilisez cet outil pour créer des tableaux croisés à partir d'autant de variables qualitatives que nécessaire, tant pour les lignes que pour les colonnes.

Dans cette section :

Description

Utilisez cet outil pour créer des tableaux croisant des variables qualitatives. Vous pouvez utiliser autant de variables qualitatives que vous voulez. Un tableau croisé est construit en croisant des variables qualitatives qui se retrouvent en lignes, et d'autres en colonnes. Les cellules d'un tableau croisé peuvent comprendre simplement le nombre d'occurrences d'un croisement donné dans le jeu de données, le pourcentage correspondant (par rapport à l'effectif total) ou une statistique calculée sur une variable quantitative.

XLSTAT permet de générer trois types de tableaux croisés : * Soit les variables qualitatives sont imbriquées les unes dans les autres (tant pour les lignes que pour les colonnes) * Soit elles sont affichées les unes après les autres (côte à côte) * Soit elles sont croisées deux à deux (Deux voies)

Exemple correspondant au premier format :

Les données originales consistent en 3 variables qualitatives (Age, Sexe, Niveau de satisfaction). L'âge et le sexe sont utilisées en variables lignes imbriqués, tandis que le niveau de satisfaction est utilisé comme variable colonne.

|Age|Sexe|1|2|3|4|5| |--|--|--|--|--|--|--| 15-25|F|27|32|40|41|44| 15-25|M|24|34|35|40|50| 26-35|F|22|28|38|44|50| 26-35|M|20|24|40|48|55| 35-60|F|19|25|30|40|44| 35-60|M|15|26|24|39|58|

Exemple correspondant au second format :

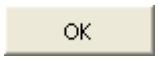
Les données originales consistent, identiques aux précédentes en 3 variables qualitatives (Age, Sexe, Niveau de satisfaction). L'âge et le sexe sont utilisés en variables lignes côte à côte, tandis que le niveau de satisfaction est utilisé comme variable colonne.

|||1|2|3|4|5| |--|--|--|--|--|--|--| Age |15-25|24|34|35|40|50| |26-35|27|32|40|41|44| |>35|20|24|40|48|55| Sexe|F|22|28|38|44|50| |M|15|26|24|39|58| ||

Statistiques :

Si par défaut un tableau croisé permet de calculer le nombre de cas correspondant à chaque croisement comme dans les exemples ci-dessus, il est possible de calculer de la même manière les % associés aux comptages (en divisant les comptages par l'effectif total et en multipliant par 100). De plus, si des variables quantitatives sont disponibles, il est possible de calculer des statistiques pour ces variables, pour chaque croisement. Par exemple dans le cas d'un tableau qui croiserait d'une part l'âge et le sexe en lignes avec la catégorie socio-professionnelle en colonne, on peut afficher dans les cellules du tableau croisé la moyenne des revenus pour les indidividus correspondant à chaque croisement possible.

Boîte de dialogue 1



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.


: cliquez sur ce bouton pour changer la façon dont vous chargez les données dans XLSTAT. Si le bouton possède une icône de souris, XLSTAT vous permet de faire une sélection à la souris des données. Si le bouton possède une icône de liste, XLSTAT vous permet de sélectionner les données à partir d'une liste. Si le bouton possède une icône orange, des boutons complémentaires avec un point d'interrogation  sont affichés pour vous permettre d'importer les données à partir de fichiers.

Données : sélectionnez les données que vous voulez faire intervenir dans la création du tableau croisé. Si des en-têtes ont été sélectionnés, vérifiez que l'option "Libellés des colonnes" est bien activée.

Poids : activez cette option si vous voulez pondérer les observations. Si vous n'activez pas cette option, les poids seront tous considérés comme valant 1. Les poids doivent être impérativement supérieurs ou égaux à 0. Si un en- tête de colonne a été sélectionné, veuillez vérifier que l'option « Libellés des colonnes » est activée.

Disposition : choisissez le type d'affichage pour les tableaux croisés. Vous pouvez choisir entre l'affichage **imbriqué** pour lequel les variables sont imbriquées dans l'ordre de sélection (l'ordre pour les variables lignes et colonnes , indépendant, est défini dans la seconde boite de dialogue), l'affichage **côte à côte** (l'ordre est aussi défini dans la seconde boîte de dialogue), ou

l'affichage **Deux voies** pour lequel XLSTAT génère tous les tableaux croisés croisant par paires uniquement les variables qualitatives lignes et colonnes sélectionnées.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

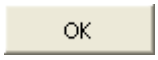
Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

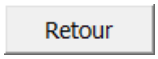
Libellés des colonnes : activez cette option si la première ligne des données sélectionnées (données, poids) contient un libellé.

Afficher l'en-tête du rapport : désactivez cette option pour que l'en-tête du rapport ne soit pas affiché.

Boîte de dialogue 2



: cliquez sur ce bouton pour lancer les calculs.



: cliquez sur ce bouton pour revenir à la boîte de dialogue précédente.

: cliquez sur ce bouton pour afficher l'aide.

Afficher les totaux : activez cette option pour afficher les statistiques marginales, pour les lignes et pour les colonnes.

Afficher les sous-totaux : activez cette option pour afficher les statistiques marginales pour chaque niveau de l'avant dernière variable en ligne.

Trier les modalités : activez cette option pour qu'au sein de chaque variable qualitative, les modalités soient triées par ordre alphabétique.

Libellés Variable-Modalité : activez cette option pour utiliser des libellés longs pour l'affichage des résultats. Les libellés Variable-Modalité sont composés du nom de la variable comme préfixe, et de la modalité du sous-échantillon comme suffixe.

Fusionner les cellules : activez cette option pour fusionner les cellules correspondant à une même variable (en lignes ou en colonnes).

Cacher les valeurs nulles : activez cette option si vous ne voulez pas afficher les valeurs nulles dans les tableaux croisés (typiquement des croisements de valeurs non rencontrés dans les données).

Un tableau par variable : activez cette option pour que si les calculs de statistiques sont demandées sur plusieurs variables quantatives, les tableaux croisés soient publiés pour chaque

variable séparément.

Un tableau par statistique : activez cette option pour que si plusieurs statistiques sont demandées, les tableaux croisés soient publiés pour chaque statistique séparément.

Variables lignes : choisissez les variables à utiliser pour les lignes du tableau croisé. Vous pouvez modifier l'ordre en sélectionnant une variable puis en utilisant les flèches vers le haut ou le bas.

Variables colonnes : choisissez les variables à utiliser pour les colonnes du tableau croisé. Vous pouvez modifier l'ordre en sélectionnant une variable puis en utilisant les flèches vers le haut ou le bas.

Calculs : Dans ce bloc vous pouvez sélectionner : * les **variables quantitatives** sur lesquelles vous voulez calculer des statistiques. Si aucune variable n'est sélectionnée seuls les comptages et/ou les % seront calculés. * les **statistiques** que vous voulez calculer pour chaque croisement. Les statistiques disponibles sont : comptables, %, données manquantes, somme des poids, somme, moyenne, médiane, écart-type. Si d'autres statistiques vous seraient utiles, contactez XLSTAT et elles seront ajoutées. Remarque : les comptages et la somme des poids sont identiques s'il n'y a pas de donnée manquante pour une variable quantitative.

Exemple

Un exemple d'utilisation de cet outil est disponible à l'adresse suivante :

http://www.xlstat.com/demo-crosstabf.htm

Tableaux croisés intelligents

Utilisez ce module pour transformer un tableau individus/variables en un tableau croisé dynamique optimisé pour la compréhension et l'analyse d'un phénomène mesuré au travers d'une variable réponse.

Dans cette section :

Description

L'outil « Tableaux croisés intelligents » est un outil unique pour créer des tableaux croisés dynamiques intelligents. Il s'appuie sur les arbres de classification en utilisant la méthode CHAID afin de faire ressortir des variables ayant un impact important sur une variable réponse donnée.

Un **tableau croisé** (ou tableau de contingence) est une représentation synthétique des occurrences observées sur une population de taille N pour des croisements des différentes catégories de deux variables.

Un **tableau croisé dynamique** permet de prendre en compte plus de deux variables et de hiérarchiser la structure du tableau. Le dynamisme du tableau provient de fonctionnalités informatiques qui permettent de naviguer dans la hiérarchie et de ne voir éventuellement que certaines classes de certaines variables.

Cet outil vous permet de construire des tableaux croisés dynamiques dont la structure est optimisée en fonction d'une variable cible. Les variables numériques continues ou discrètes explicatives (celles dont les catégories constituent les lignes et les colonnes du tableau) sont automatiquement découpées en des classes qui permettent d'optimiser la qualité du tableau.

La variable cible peut être une variable qualitative ou une variable quantitative.

L'outil « Tableaux croisés intelligents » se base sur la méthode des arbres de classification afin, d'une part, de discrétiser les variables quantitatives et, d'autre part, d'identifier les variables ayant le plus fort impact sur la variable réponse qu'elle soit qualitative ou quantitative (voir chapitre de l'aide sur les arbres de classification). La méthode CHAID est privilégiée car elle s'adapte bien à la problématique de la représentation d'un tableau croisé dynamique.

Cet outil vous permet aussi bien d'utiliser des variables qualitatives que quantitatives en les discrétisant de manière optimale (basé sur l'algorithme CHAID).

Lorsque vous utilisez cette fonction, vous verrez une boîte de dialogue vous permettant de sélectionner les variables que vous voulez utiliser dans le tableau croisé dynamique. Afin de vous aider un indice de qualité d'explication correspondant à chacune des variables est affiché (voir plus bas pour la description de cet indice).

Vous pouvez régler plusieurs paramètres. Néanmoins l'utilisation des paramètres par défaut doit donner les meilleurs résultats. Il est donc possible de choisir le type de discrétisation des variables quantitatives. La méthode automatique étant la discrétisation à l'intérieur de l'algorithme CHAID. L'outil offre aussi la possibilité de régler un indice de sensibilité pour la construction de l'arbre de classification (voir plus bas pour la description de cet indice).

Indice du score des variables explicatives

Pour évaluer la contribution des variables explicatives, un indice de score associé à chaque variable a été mis en place. Cet indice sera différent suivant que la variable réponse est quantitative ou qualitative.

Dans le cas d'une **variable réponse quantitative**, l'indice du score est l'importance de la variable suivant la définition de Breiman et al. (1984). Celle-ci est définie par :

$$Score(var_i) = \sum_{j \in T} i^2 I(var_i \in Node_j)$$

Avec $i^2 = \frac{w_i \times w_j}{w_i + w_j}(\bar{y_i} - \bar{y_j})^2$, avec i et j nœud fils du nœud j étudié et

$$I(var_i \in Node_j) = \begin{cases} 1 \text{ si le noeud j concerne la variable i} \\ 0 \text{ sinon} \end{cases} .$$

Les poids w sont calculés avec : $w_i = \frac{n_i}{N}(1 - \frac{n_i}{N})$, n_i étant le nombre d'observations associées à la feuille et N le nombre d'observations associées au nœud parent.

Dans le cas d'une **variable réponse qualitative**, XLSTAT demande à l'utilisateur de sélectionner une modalité cible qui sera prise en compte dans le tableau croisé ainsi que dans le calcul du score. L'indice de score associé à chaque variable est alors :

$$Score(var_i) = \sum_{j \in T} i^2 I(var_i \in Node_j)$$

Avec $i^2 = \frac{w_i \times w_j}{w_i + w_j} \sum_{k=1}^{nb_{mod}} (p_{ik} - p_{jk})^2$ avec i et j nœud fils du nœud étudié et

$$I(var_i \in Node_j) = \begin{cases} 1 \text{ si le noeud j concerne la variable i} \\ 0 \text{ sinon} \end{cases}$$

Les poids w sont calculés avec : $w_i = \frac{n_i}{N}(1 - \frac{n_i}{N})$

n_i étant le nombre d'observations associées à la feuille et N le nombre d'observations associées au nœud parent. Les probabilités sont les probabilités associées à chacune des modalités de la variable réponse pour chacune des feuilles.

Indice de sensibilité de la construction de l'arbre

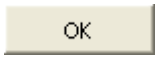
La construction d'un arbre de classification nécessite de fixer un certain nombre de paramètres (la profondeur maximale, la taille des feuilles, les seuils de regroupement et de séparation…). Afin de simplifier l'utilisation de cet outil, un indice de sensibilité a été mis au point. Celui-ci prend des valeurs entre 0 et 1.

Lorsque cet indice est proche de 0, alors la construction de l'arbre est peu sensible à de petites différences. Le nombre d'intervalles dans la discrétisation des variables explicatives quantitatives sera plus faible et la taille de l'arbre petite. Ce sont donc les contributions les plus fortes qui seront révélées par le tableau croisé dynamique.

Lorsque cet indice est proche de 1, alors la construction de l'arbre est très sensible à de petites différences. Le nombre d'intervalles dans la discrétisation des variables explicatives quantitatives sera plus grand et la taille de l'arbre grande. Toutes les contributions seront révélées par le tableau croisé dynamique.

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.

Onglet **Général**:

Y / Variable réponse : Sélectionnez la variable que vous voulez modéliser. Si un en-tête a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

Choisissez le format de la variable réponse que vous avez sélectionnée :

- Quantitative : si vous choisissez cette option, vous devez sélectionner une variable quantitative.
- Qualitative : si vous choisissez cette option, vous devez sélectionner une variable qualitative. Vous devrez alors sélectionner une modalité cible dans la liste défilante qui apparaît sur la droite.

X / Variables explicatives

Quantitatives : activez cette option si vous voulez utiliser des variables explicatives quantitatives. Ensuite, sélectionnez une ou plusieurs variables explicatives quantitatives. Si des en-têtes ont été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

Qualitatives : activez cette option si vous voulez utiliser des variables explicatives qualitatives. Ensuite, sélectionnez une ou plusieurs variables explicatives qualitatives. Si des en-têtes ont été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

Poids des observations : activez cette option si vous voulez pondérer les observations. Si vous n'activez pas cette option, les poids seront tous considérés comme valant 1. Les poids doivent être impérativement supérieurs ou égaux à 0. Si un en-tête de colonne a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des variables : activez cette option si la première ligne des données sélectionnées (données et libellés des observations) contient un libellé.

Libellés des observations : activez cette option si vous voulez utiliser des libellés d'observations disponibles sur une feuille Excel pour l'affichage des résultats. Si l'option « Libellés des variables » est activée, la première cellule de la sélection doit comprendre un en-tête. Si vous n'activez pas cette option, des libellés seront automatiquement créés (Obs1, Obs2, …).

Onglet **Options** :

Sensibilité : vous pouvez modifier la valeur de la sensibilité. Plus on s'approche de 1, plus l'arbre de classification est grand, plus on s'approche de 0, moins il sera grand. Pour une description de cet indice, voir la partie description de ce chapitre. La valeur par défaut est de 0,5.

Discrétisation des X : cette option est active uniquement si des variables explicatives quantitatives ont été sélectionnées.

- **Automatique** : activez cette option pour effectuer une discrétisation automatique en utilisant les méthodes associées aux arbres de classification. C'est l'option par défaut et la plus efficace.
- **Amplitude égale** : activez cette option pour discrétiser les variables quantitatives de manière à avoir des intervalles d'amplitude égale. Le nombre d'intervalle doit aussi être donné.
- **Fréquence égale** : activez cette option pour discrétiser les variables quantitatives de manière à avoir des intervalles de fréquence égale. Le nombre d'intervalle doit aussi être donné.
- **Défini par l'utilisateur** : activez cette option pour discrétiser les variables quantitatives en utilisant des intervalles définis par l'utilisateur. Il faudra donc sélectionner un tableau de données comportant pour chaque variable une colonne avec toutes les bornes des intervalles.

Onglet **Données manquantes** :

Supprimer les observations : activez cette option pour supprimer les observations comportant des données manquantes.

Estimer les données manquantes : activez cette option pour estimer les données manquantes avant le début des calculs.

- **Moyenne ou mode** : activez cette option pour estimer les données manquantes en utilisant la moyenne (variables quantitatives) ou le mode (variables qualitatives) pour les variables correspondantes.
- **Plus proche voisin** : activez cette option pour estimer les données manquantes d'une observation en recherchant le plus proche voisin de l'observation.

Onglet **Sorties**:

Statistiques descriptives : activez cette option pour afficher le tableau des statistiques descriptives pour les variables sélectionnées.

Discrétisation : activez cette option pour afficher le tableau récapitulatif des variables discrétisées.

Contributions : activez cette option pour afficher le tableau des contributions et le diagramme en bâtons correspondant.

Tableau croisé : activez cette option pour afficher le tableau croisé dynamique.

Résultats

Le premier tableau affiché donne les statistiques descriptives associées aux variables sélectionnées en fonction de leur type.

Le second tableau récapitule les discrétisations utilisées pour les variables sélectionnées.

Le troisième tableau donne les contributions des variables (contribution brute, relative en %, et cumulée). Il permet de détecter rapidement quelles sont les variables qui ont le plus d'impact sur la variable cible. Un diagramme en bâtons correspondant aux contributions est aussi affiché.

Le principal résultat fourni est le tableau croisé dynamique. Chaque case du tableau correspond à une combinaison unique de valeurs des variables explicatives et est décrite par 4 valeurs qui peuvent être affichées ou non en fonction des préférences de l'utilisateur :

- **Moyenne cible** : c'est la moyenne de la variable cible sur la sous-population correspondant à la combinaison dans le cas d'une variable continue et le pourcentage d'occurrence de la modalité cible de la variable réponse lorsque celle-ci est qualitative ;
- **Taille cible** : comptage des occurrences de la modalité cible de la variable réponse dans le cas d'une variable binaire ;
- **Taille population** % : pourcentage de la population totale qui correspondant à la combinaison ;
- **Taille Population** : effectif de la population correspondant à la combinaison.

Exemple

Un exemple portant sur des données d'un recensement effectué aux Etats-Unis est disponible en permanence sur le Centre d'aide XLSTAT :

http://www.xlstat.com/demo-pivotf.htm

Bibliographie

Breiman, L., Friedman, J.H., Olshen, R. A. and Stone, C.J. (1984). Classification and regression tree, Chapman & Hall.

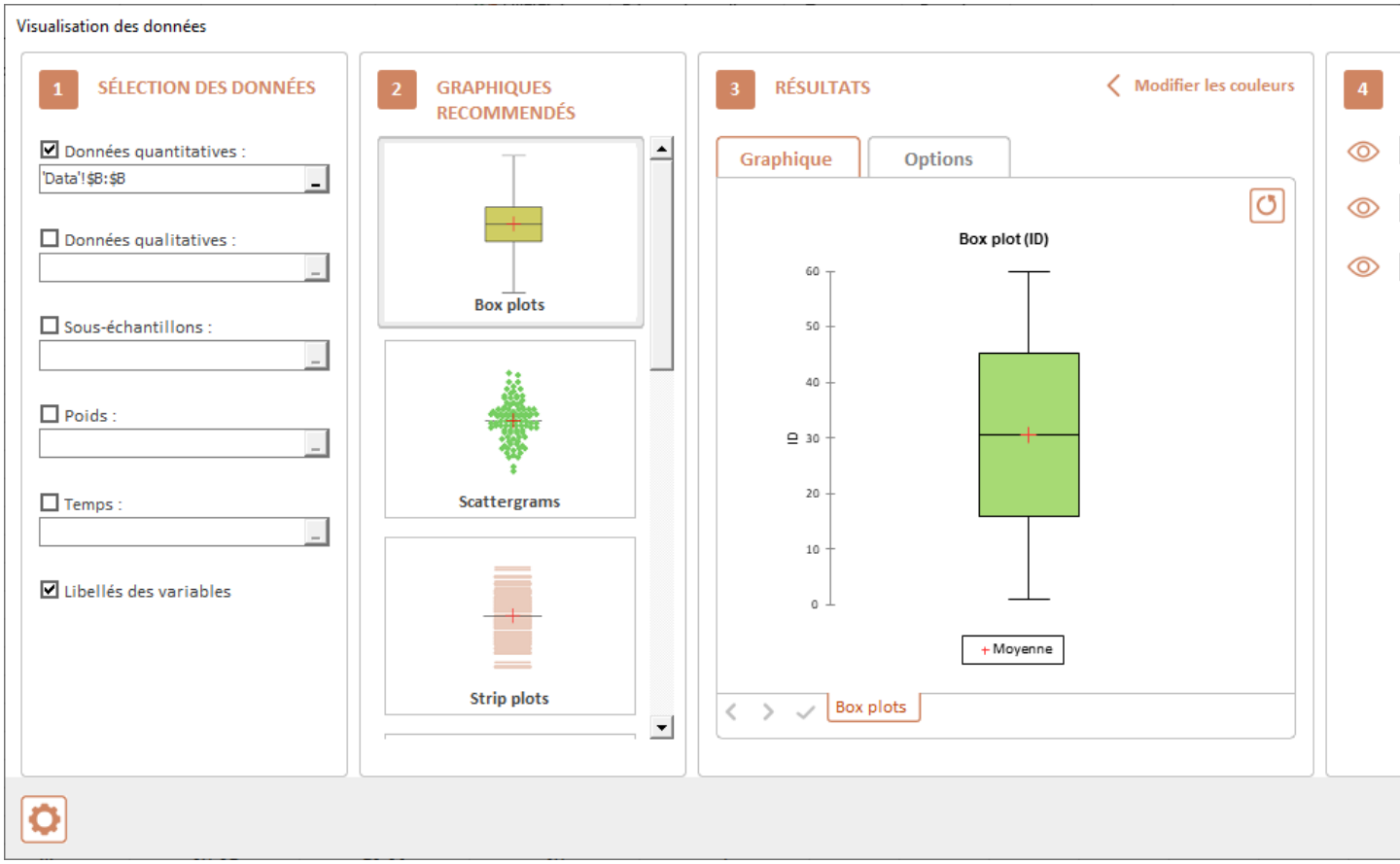
Visualisation des données

DataViz

Utilisez le DataViz pour trouver en quelques clics le graphique idéal adapté à vos besoins et personnalisez-le en toute simplicité.

Dans cette section :

Générer un graphique personnalisable en quelques clics



Le DataViz est l'outil parfait pour sélectionner un graphique qui correspond à ses données sans avoir à naviguer entre plusieurs fonctionnalités.

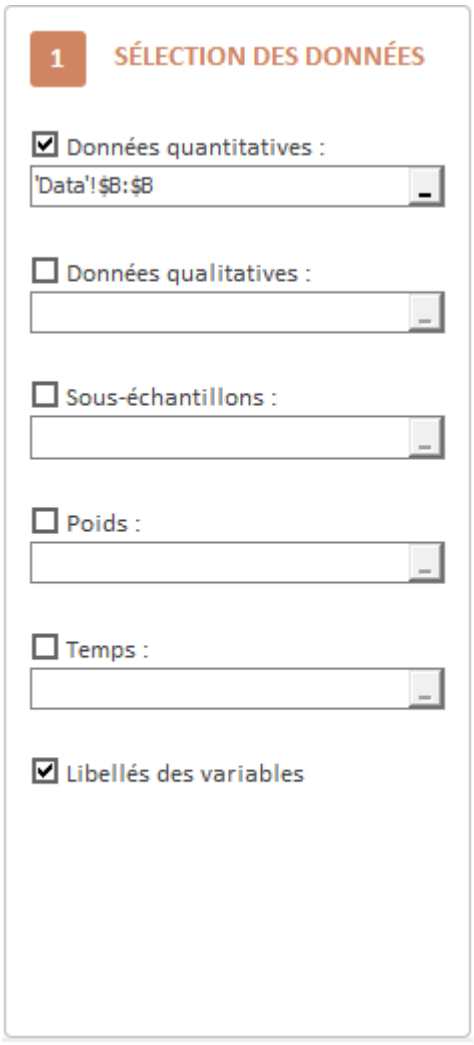
Grâce au DataViz, générez des graphiques en seulement 4 étapes.

Le graphique choisi peut être personnalisé et exporté dans Excel pour vous permettre de l'intégrer à vos analyses et rapports.

Explorez les riches fonctionnalités du DataViz

1. Sélectionner les données à visualiser

Sélectionner les données à visualiser



L'outil Dataviz permet de sélectionner tout type de données afin de générer une visualisation appropriée :

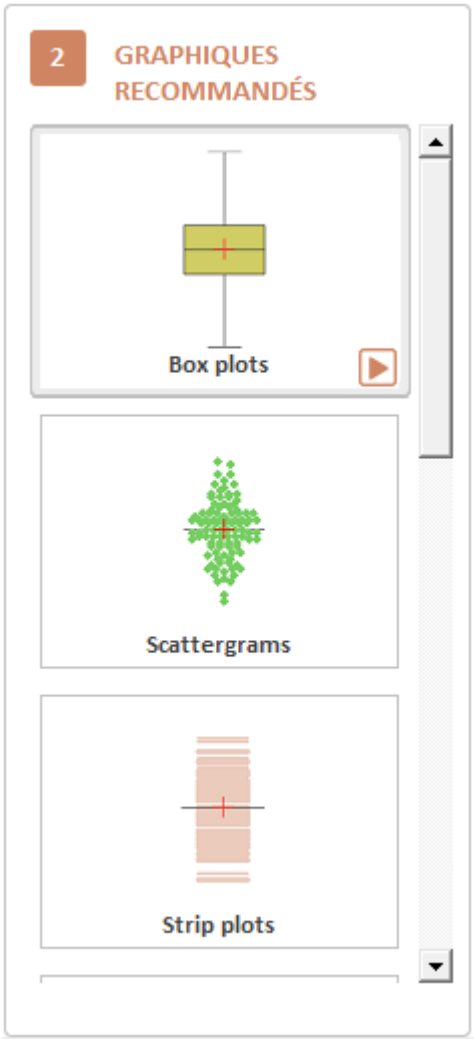
- données quantitatives ;
- données qualitatives ;
- sous-échantillons ;
- poids ;
- temps.

Une fonction de détection automatique vous avertit en cas d'incompatibilité entre les données sélectionnées et le type de données choisi.

Indiquez si les données fournies comportent un en-tête.

2. Choisir le graphique qui illustre au mieux les données

Liste des graphiques recommandés



Une liste de graphiques recommandés est affichée et montre toutes les possibilités de visualisation à disposition.

La miniature de chaque type de graphique facilite la sélection de celui qui mettra en évidence le résultat désiré de manière intuitive.

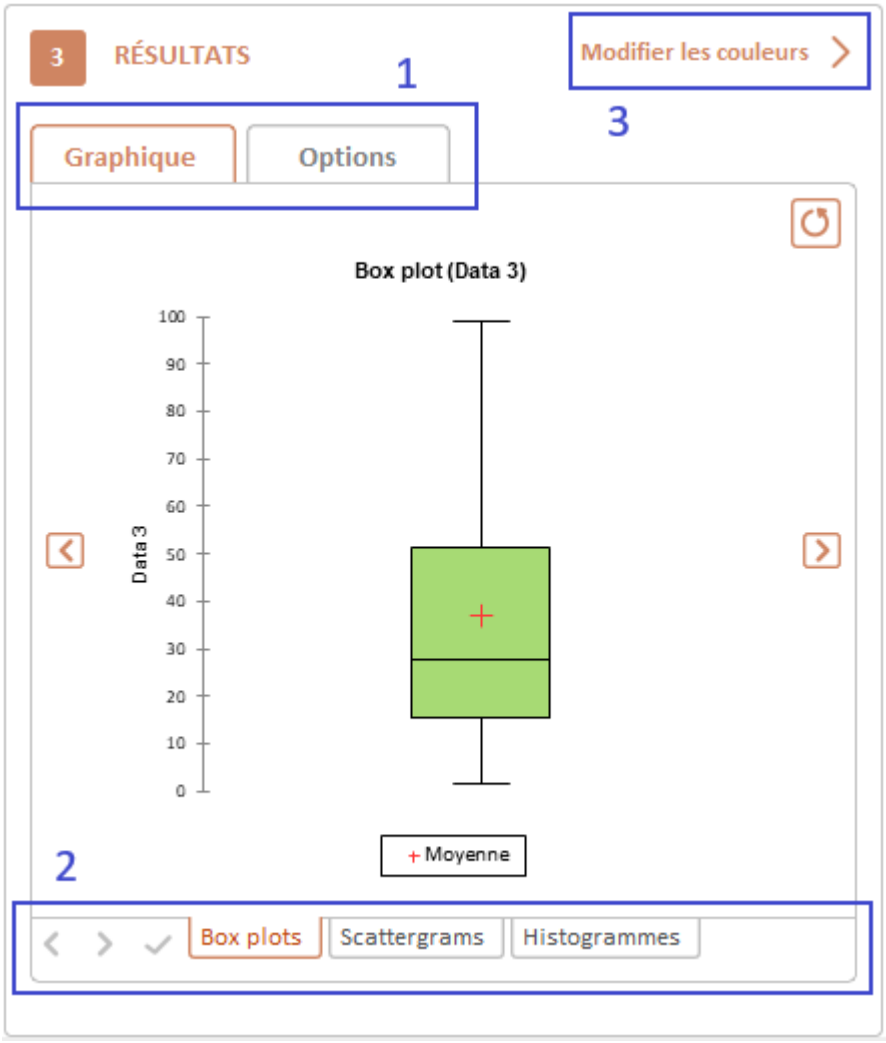
Profitez d'une mise à jour très rapide de la liste des graphiques, après un changement dans la sélection des données.

Cliquer sur une fois le type de graphique choisi, afin de générer le graphique.

Note : étant encore en version bêta, le DataViz pourrait ne pas proposer de graphiques pour toutes les combinaisons de types de données. Cependant, de nouveaux graphiques seront progressivement intégrés dans les versions à venir.

3. Paramétrer le graphique

Paramétrer le graphique



Une fois le graphique généré, le DataViz offre une gamme d'actions dynamiques.

Naviguez entre les différents graphiques générés grâce aux chevrons .

Cette fonctionnalité prend son sens lorsqu'un graphique par catégorie est créé, par exemple.

Retrouvez le graphique initial, avant toute modification d'options ou de couleurs, en cliquant tout simplement sur .

En présence d'une variable temporelle (champ **Temps**), interagissez avec le graphique grâce au panel de boutons .

Ces boutons permettent, dans l'ordre, de :

- reculer dans le temps pour saisir l'évolution du graphique ;
- mettre en pause la lecture ;
- faire une lecture de l'évolution du graphique dans le temps ;
- accélérer ou ralentir la vitesse de lecture.

Le simple clic sur le titre d'un graphique ou le titre d'un de ses axes vous permet de les

modifier.

Visualisez un graphique ou redéfinissez les paramètres spécifiques à chaque graphique en naviguant entre les onglets **Graphique** et **Options** (*Zone 1*). Expérimentez et régénérez les paramètres à votre guise pour des résultats à votre image.

Explorez l'historique des graphiques générés avec aisance en utilisant les chevrons (*Zone 2*). Faites glisser les onglets à gauche ou à droite lorsque les onglets ne sont plus visibles, ou retournez instantanément à l'onglet en cours en cliquant sur (*Zone 2*).

Retrouvez dans chaque onglet de l'historique un graphique généré pour un ensemble « données sélectionnées - graphique recommandé ». L'onglet actif sera coloré en orange.

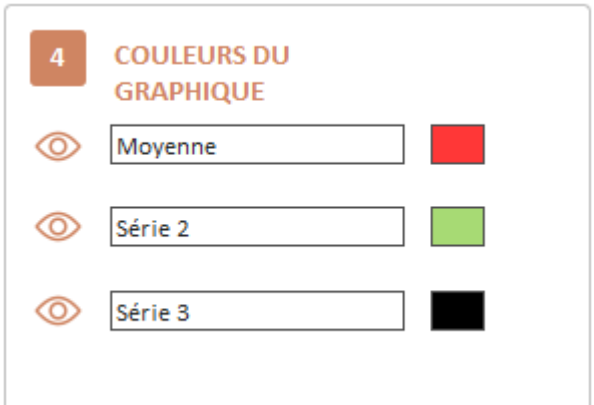
Un seul clic sur le nom d'un onglet permet sa modification. Pour le retirer, un clic droit sur son nom suffit.

Enfin, personnalisez les couleurs du graphique en cliquant sur Modifier les couleurs (*Zone 3*).

Note : Modifier les options ne créera pas un nouvel onglet.

4. Personnaliser les couleurs du graphique

Cliquer sur Modifier les couleurs pour visualiser et personnaliser les séries présentes dans vos graphiques.



Personnaliser les couleurs

L'outil Dataviz permet de sélectionner tout type de données afin de générer une visualisation appropriée : Chaque série est représentée par une ligne, offrant trois options interactives :

- obtenez un aperçu immédiat des séries dans le graphique généré en un simple clic sur ;
- renommez aisément une série via la zone de texte respective, sans aucune incidence sur le graphique. Cela facilite grandement l'identification de chaque série, surtout lorsqu'un grand nombre est utilisé ;
- exprimez votre créativité en choisissant une ou plusieurs nouvelles couleurs, en cliquant sur le carré de couleur. Grâce au sélecteur de couleurs, un vaste éventail de teintes est à votre disposition.

5. Exporter le graphique personnalisé pour une utilisation optimale dans Excel

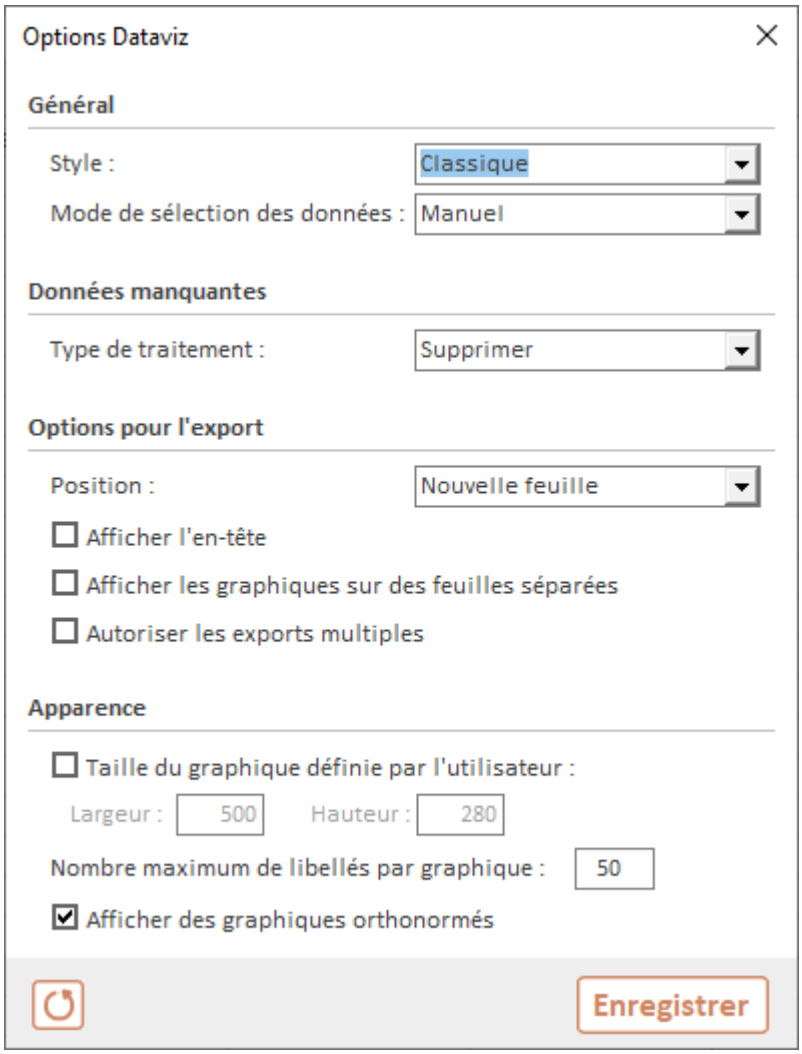
Une fois votre graphique parfaitement élaboré, il ne reste qu'à l'exporter.

Pour cela, il vous suffit de cliquer sur l'icône 

située dans le coin inférieur droit de la fenêtre, permettant ainsi d'exporter le ou les graphiques figurant dans l'onglet en cours.

Pour plus de détails, veuillez consulter la section Boîte de dialogue des options générales.

Boîte de dialogue des options générales



- **Général** :
    - **Style** : cette option permet de modifier les couleurs du graphique généré. Choisissez le style que vous préférez parmi « Classique » qui correspond au format historique de XLSTAT, « Moderne » qui correspond à un autre jeu de couleurs et « Scientifique » qui n'utilise que du noir, du blanc et des gris. Cette option ne s'applique pas à tous les graphiques.

- **Mode de sélection des données** :
    - **Manuel** : choisissez cette option pour sélectionner directement les données dans une feuille Excel.
    - **Variables** : choisissez cette option pour qu'une boîte de dialogue s'affiche avec la liste des données automatiquement repérées dans la feuille active. Vous n'aurez plus qu'à cocher les données à utiliser comme entrée.

- **Données manquantes** :
    - **Type de traitement** : choisissez la manière dont les données manquantes doivent être traités.
        - **Refuser** : choisissez cette option pour ne pas générer de graphique lorsqu'il y a des données manquantes.
        - **Supprimer** : choisissez cette option pour supprimer les données manquantes avant de générer un graphique.
        - **Remplacer** : choisissez cette option pour remplacer les données manquantes par leur moyenne avant de générer un graphique.

  Note : attention, certains graphiques n'ont pas encore la gestion des données manquantes. Ce paramètre ne s'applique donc pas à eux.

- **Options pour l'export**
    - **Position** :
        - **Nouvelle feuille** : choisissez cette option pour que les graphiques soient exportés dans une nouvelle feuille du classeur actif.
        - **Nouveau classeur** : choisissez cette option pour que les graphiques soient exportés dans un nouveau classeur.
        - **Emplacement personnalisé** : choisissez cette option pour pouvoir sélectionner une cellule qui correspondra à l'emplacement de l'export.
    - **Afficher l'en-tête** : activez cette option pour afficher un en-tête dans les feuilles d'export. L'en-tête comprend des boutons pour relancer le graphique exporté dans le DataViz et pour exporter les graphiques présents dans la feuille dans Word ou Powerpoint.
    - **Afficher les graphiques sur des feuilles séparées** : activez cette option pour que les graphiques soient exportés sur des feuilles graphiques séparées.

  Note : lorsqu'un graphique est affiché sur une feuille Excel standard, vous pouvez le convertir en feuille graphique séparée, en suivant ces étapes : - sélectionner le graphique ; - faire un clic droit sur « Déplacer le graphique » - choisir « Nouvelle feuille ».

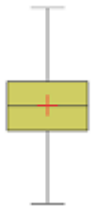
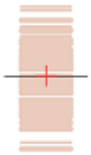
- **Autoriser les exports multiples** : activez cette option si vous souhaitez pouvoir exporter plusieurs graphiques avant de fermer la boîte de dialogue du DataViz.

- **Apparence** :

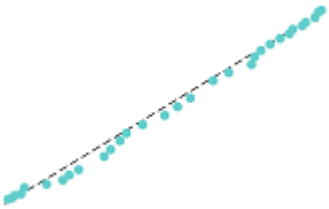
- **Taille du graphique définie par l'utilisateur** : activez cette option pour que XLSTAT affiche des graphiques dont la taille est exactement définie par les valeurs ci-dessous.
- **Largeur** : entrez la valeur en points de la largeur des graphiques.
- **Hauteur** : entrez la valeur en points de la hauteur des graphiques.
- **Nombre maximum de libellés par graphique** : entrez le nombre maximum de libellés à afficher sur un graphique.
- **Afficher des graphiques orthonormés** : activez cette option pour que les graphiques issus d'analyses factorielles soient orthonormés. Cela permet d'avoir automatiquement des échelles identiques pour les abscisses et les ordonnées, et d'éviter des interprétations erronées du fait d'effets de dilatation artificiels.

Liste des graphiques recommandés

Graphiques recommandés pour les données quantitatives

Box plot	**Strip plot**	**Scattergram**
		
Idéal pour les **données quantitatives**, vous pouvez inclure des **sous-échantillons** et ajouter des **poids** aux données. Appelé aussi « boîte à moustache », il s'agit d'un rectangle affichant le minimum, le 1er quartile, la médiane, la moyenne, le 3ème quartile, ainsi que les deux limites au-delà desquelles on peut considérer que les valeurs sont anormales. La moyenne est affichée sous la forme d'un + rouge, et la médiane sous la forme d'une ligne noire.	Idéal pour les **données quantitatives**, vous pouvez inclure des **sous-échantillons** et ajouter des **poids** aux données. Représente sous forme de bandes (*strip* en anglais) les données de l'échantillon. Sur un intervalle donné, plus les bandes sont serrées ou épaisses plus il y a de données.	Idéal pour les **données quantitatives**, vous pouvez inclure des **sous-échantillons** et ajouter des **poids** aux données. Donne une idée de la distribution et de la pluralité éventuelle des modes d'un échantillon. Tous les points sont représentés, ainsi que la moyenne et la médiane.

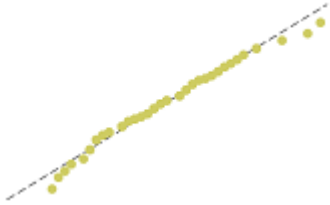
Graphique P-P (loi normale)



Idéal pour les **données quantitatives**, vous pouvez inclure des **sous-échantillons** et ajouter des **poids** aux données.

Permet de comparer la fonction de répartition empirique d'un échantillon à celle d'un échantillon distribué suivant une loi normale de même moyenne et même variance. Si l'échantillon suit une loi normale, les points doivent être confondus avec la première bissectrice du plan.

Graphique Q-Q (loi normale)



Idéal pour les **données quantitatives**, vous pouvez inclure des **sous-échantillons** et ajouter des **poids** aux données.

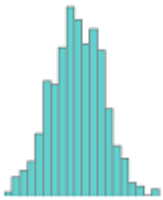
Permet de comparer les quantiles de l'échantillon à ceux d'un échantillon distribué suivant une loi normale de même moyenne et même variance. Si l'échantillon suit une loi normale, les points doivent être confondus avec la première bissectrice du plan.

Graphique des moyennes

Idéal pour les **données quantitatives**, vous pouvez inclure des **sous-échantillons** et ajouter des **poids** aux données.

Représente les moyennes de chacune des variables, sous forme de diagrammes en bâtons. Il est aussi possible d'afficher les **barres d'erreurs** sous trois formes différentes.

Histogramme



Idéal pour les **données quantitatives**, vous pouvez inclure des **sous-échantillons** et ajouter des **poids** aux données.

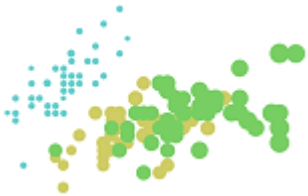
Permet d'avoir rapidement une idée de la distribution d'un échantillon de données quantitatives continues ou discrètes.

Tests de corrélation

Idéal pour les **données quantitatives**, il nécessite au moins **2 données qualitatives**. Vous pouvez inclure des **sous-échantillons** et ajouter des **poids** aux données.

Représente une matrice de corrélation dont les valeurs ont été colorées selon une échelle personnalisable. Cette carte permet de voir en un coup d'œil les variables fortement corrélées.

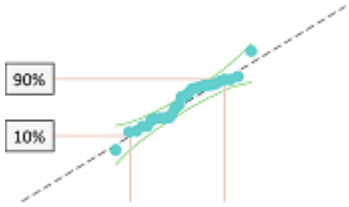
Nuage de points



Idéal pour les **données quantitatives**, vous pouvez inclure des **sous-échantillons** et ajouter des **poids** aux données.

Représente les données en 2 ou 3 dimensions.

Diagramme de probabilité



Idéal pour les **données quantitatives**, vous pouvez inclure des **sous-échantillons** et ajouter des **poids** aux données.

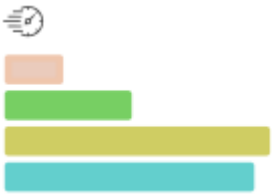
Aide à vérifier visuellement si un échantillon provient d'une population suivant une distribution donnée.

Diagramme en tornade

Idéal pour une variable contenant des **données quantitatives** accompagnée de **sous-échantillons** ou pour **2 données quantitatives**.

Similaire à un graphique en barres, il permet de comparer l'importance relative de deux variables.

Bar chart race



Idéal pour modéliser des **données quantitatives** accompagnées d'une donnée de **temps**, vous pouvez inclure des **sous-échantillons**.

Le bar chart race (ou course de bar chart) permet de visualiser l'évolution d'une variable au cours du temps sur un seul graphique dynamique.

Motion Chart

Idéal pour modéliser **2 données quantitatives** accompagnées d'une donnée de **temps**. Vous pouvez inclure des **sous-échantillons**.

Permet de visualiser l'évolution de plusieurs variables (jusqu'à 3), mesurées sur plusieurs individus, au cours du temps sur un seul graphique dynamique.

Graphiques recommandés pour les données qualitatives

Diagramme en secteurs

Idéal pour les **données qualitatives**, vous pouvez inclure des **sous-échantillons** et ajouter des **poids** aux données.

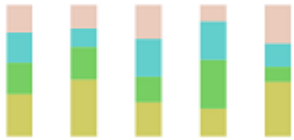
Représente sous forme de diagramme en secteurs (ou camemberts) les effectifs ou les fréquences des différentes modalités des variables qualitatives.

Diagramme en bâtons

Idéal pour les **données qualitatives**, vous pouvez inclure des **sous-échantillons** et ajouter des **poids** aux données.

Représente sous forme de diagrammes en bâtons, les effectifs ou les fréquences des différentes modalités des variables qualitatives.

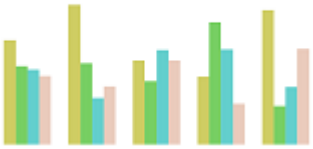
Barres empilées



Adapté pour les **données qualitatives** subdivisées en **sous-échantillons**. Vous pouvez inclure des **poids** aux données.

Permet de comparer les effectifs ou les fréquences des sous-échantillons à ceux d'un échantillon complet.

Barres multiples



Adapté pour les **données qualitatives** subdivisées en **sous-échantillons**. Vous pouvez inclure des **poids** aux données.

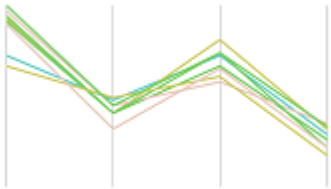
Permet de comparer les effectifs ou les fréquences des sous-échantillons à ceux d'un échantillon complet.

Graphique 2D pour tableaux croisés

Idéal pour des **données qualitatives**, il nécessite au moins **2 données qualitatives** pour être exploité.

Permet de générer un graphique 2D montrant l'importance relative des diverses combinaisons que vous pouvez obtenir lorsque vous créez un tableau de contingence ou plus généralement un tableau croisé.

Graphiques recommandés pour des données mixtes

Diagramme en radar	Graphique en coordonnées parallèles	Graphique sémantique différentiel
		
Adapté aux **données quantitatives**, il nécessite **une donnée qualitative** pour être exploité. Permet d'évaluer différents choix en fonction de plusieurs variables.	Adapté aux **données quantitatives** et **qualitatives**, vous pouvez inclure des **sous-échantillons** et ajouter des **poids** aux données. Permet de visualiser des données multidimensionnelles sur un même graphique à deux dimensions.	Adapté aux **données quantitatives** et **qualitatives**. Permet de visualiser les notes attribuées par des sujets à des objets pour différents critères.

Diagrammes de probabilités

Utilisez cet outil pour créer des diagrammes de probabilité et vérifier visuellement si un échantillon provient d'une population suivant une distribution donnée.

Dans cette section :

Description

Les diagrammes de probabilité sont une méthode ancienne (Hazen, 1914), qui a été peu à peu perfectionnée et largement utilisée, notamment au travers de l'utilisation de papier de probabilité. Ils sont particulièrement utiles pour contrôler visuellement si un échantillon suit une distribution donnée.

Toutes les distributions disponibles dans XLSTAT peuvent être utilisées (voir l'outil Histogrammes pour obtenir la liste). L'ajustement de la distribution peut êtte fait par XLSTAT avant l'affichage du diagramme. La distribution qui s'ajuste le mieux peut être automatiquement identifiée par XLSTAT, ou vous pouvez choisir une distribution spécifique et laisser XLSTAT faire l'ajustement, ou vous pouvez entrer la valeur des paramètres de la distribution directement.

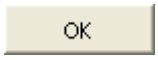
Soient $\{x_1, x_2, \ldots, x_n\}$ les statistiques d'ordre pour un échantillon de taille n censé suivre la distribution $F(x)$. Pour construire un diagramme de probabilité, les points d'abscisse x_i et d'ordonnée $F^{-1}(p_i)$ sont affichés, où p_i est l'estimateur de $F(x_i)$, appelé position. Plusieurs approches ont été proposées pour calculer p_i. XLSTAT propose les options suivantes :

- Blom (1958): $p_i = (i - 0.375)/(n + 0.25)$,
- Hazen (1914): $p_i = (i - 0.5)/n$,
- Weibull (1939): $p_i = i/(n + 1)$,
- Filliben (1975): $p_1 = 1 - 0.5^n/p_n = 0.5^n/p_i = (i - 0.3175)/(n + 0.365)$ $(1 < i < n)$.

XLSTAT offre également la possibilité de calculer l'estimateur des statistiques d'ordre pour une distribution donnée au travers de simulations Monte Carlo. Les simulations Monte Carlo sont aussi utilisées par XLSTAT pour calculer des intervalles de confiance. Des intervalles de confiance asymptotiques sont aussi proposés (Chambers, 1983).

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.

Onglet **Général**:

Données : sélectionnez des données quantitatives. Si plusieurs échantillons sont sélectionnés, XLSTAT fera les calculs pour chacun des échantillons indépendamment. Si des en-têtes ont été sélectionnés, veuillez vérifier que l'option « Libellés des échantillons » est activée.

Poids : activez cette option si vous voulez pondérer les observations. Si vous n'activez pas cette option, les poids seront tous considérés comme valant 1. Les poids doivent être impérativement supérieurs ou égaux à 0. Si un en-tête de colonne a été sélectionné, veuillez vérifier que l'option « Libellés des échantillons » est activée.

Sous-échantillons : Check this option to select a column showing the names or indexes of the sub-samples for each of the observations.

- **Couleurs uniquement** : activez cette option pour utiliser l'information uniquement pour la coloration des points sur le diagramme, sinon un diagramme sera produit pour chaque sous échantillon.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des échantillons : activez cette option si la première ligne des données sélectionnées (données, sous échantillons, poids) contient un libellé.

Onglet **Options** :

Distribution : choisissez la distribution qui doit être utilisée pour l'ajustement. Voir la partie description de l'ajustement à une loi de probabilité pour plus d'information sur les lois proposées. L'option **automatique** vous permet de laisser XLSTAT identifier la distribution la plus adéquate sur la base d'un test de Kolmogorov.

Paramètres : vous pouvez choisir d'**entrer** d'entrer les paramètres de la loi, ou de les **estimer**. Si vous choisissez d'entrer les paramètres, vous devez entrer la valeur des paramètres.

Méthode d'estimation : choisissez la méthode d'estimation des paramètres de la distribution choisie (voir la section description pour plus de détails sur les méthodes d'estimation).

- **Moments** : activez cette option pour utiliser la méthode des moments.
- **Maximum de vraisemblance** : activez cette option pour utiliser la méthode du maximum de vraisemblance. Vous pouvez alors modifier la valeur limite de **convergence** qui, une fois atteinte, permet de considérer que l'algorithme a convergé. Valeur par défaut : 0,00001.

Méthode : choisissez la méthod que vous voulez utiliser pour calculer les positions (voir la section description pour plus d'information). Si la méthode choisie est Monte Carlo, vous pouvez choisir le nombre de simulations et le temps maximum time à passer sur les simulations.

Onglet **Données manquantes** :

Ne pas accepter les valeurs manquantes : activez cette option pour que XLSTAT empêche la poursuite des calculs si des valeurs manquantes sont détectées.

Supprimer les observations :

- **Pour l'échantillon correspondant** : activez cette option pour ne pas prendre en compte une observation dont l'une des données est manquante, uniquement pour les échantillons pour lesquels une donnée est manquante.
- **Pour tous les échantillons** : activez cette option pour ne pas prendre en compte une observation dont l'une des données est manquante, pour tous les échantillons sélectionné.

Estimer les données manquantes : activez cette option pour estimer les données manquantes en utilisant la moyenne de l'échantillon.

- **Moyenne ou mode** : activez cette option pour estimer les données manquantes en utilisant la moyenne (variables quantitatives) ou le mode (variables qualitatives) pour les variables correspondantes.

Onglet **Sorties** :

Statistiques descriptives : activez cette option pour afficher les statistiques descriptives pour les échantillons sélectionnés.

Onglet **Graphiques** :

Echelle Log : activez cette option pour utiliser une échelle log sur le diagramme de probabilité.

Abscisses=Observées : activez cette option pour que les valeurs observées correspondent à l'axe des abscisses. Sinon elles correspondent aux ordonnées.

Afficher les % : activez cette option pour utiliser les % cumulés comme échelle pour l'axe utilisé pour les positions.

Intervalles de confiance : activez cette option pour afficher des intervalles de confiance. La valeur que vous entrez (comprise entre 1 et 99), en **pourcentage**, est utilisée pour déterminer les intervalles de confiance sur les estimations. La valeur par défaut est 95.

Percentiles : activez cette option et sélectionnez jusqu'à quatre percentiles que vous voulez afficher sur le diagramme.

Quantiles : activez cette option et sélectionnez jusqu'à quatre quantiles que vous voulez afficher sur le diagramme.

Results

Statistiques descriptives : le tableau de statistiques descriptives présente pour toutes les variables sélectionnées des statistiques simples. Sont affichés le nombre d'observations, le nombre de données manquantes, le nombre de données non manquantes, la moyenne, et l'écart-type (non biaisé).

Si une distribution a été ajustée aux données, les résultats de l'ajustement et les paramètres estimés de la distribution sont affichés.

Le diagramme de probabilité est ensuite affiché, avec, si l'option correspondante a été sélectionnée dans la boîte de dialogue, les intervalles de confiance.

Exemple

Un exemple montrant comment générer un diagramme de probabilité avec XLSTAT est disponible sur :

http://www.xlstat.com/demo-probaplotsf.htm

References

Blom G. (1958). Statistical Estimates and Transformed Beta Variables. John Wiley, New York.

Chambers J.M., Cleveland W.S., Kleiner B. and Tukey P.A. (1983). Graphical Methods for Data Analysis. Duxbury, Boston.

Cunname C. (1978). Unbiased plotting positions - A review. Journal of Hydrology, **37**, 205-222.

Filliben J.J. (1975). The probability plot correlation coefficient test for normality. Technometrics, **17**, 111-117

Hazen A. (1914). Storage to be provided in the impounding reservoirs for municipal water supply. Transactions of the American Society of Civil Engineers, 77, 1547-1550.

Kimball B.F. (1960). On the choice of plotting positions on probability paper. Journal of the American Statistical Association, **55**, 546-560.

Looney S.W. and Gulledge T.R. (1985). Use of the correlation coefficient with normal probability plots. Am. Stat., 39(1), 75-79.

Royston J. P. (1982). Algorithm AS 177: Expected normal order statistics (exact and approximate). *Journal of the Royal Statistical Society. Series C*, **31(2)**, 161-165.

Weibull W. (1939). The phenomenon of rupture in solids. Ingeniors Vetenskaps Akademien Handingar, **153**, 17.

Nuages de points

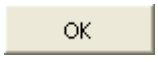
Utilisez cet outil pour créer des graphiques en 2 dimensions ou en 3 dimensions (la 3-ème dimension étant représentée par la taille du point), voire en 4 dimensions (une variable qualitative peut être sélectionnée). Cet outil permet aussi la création de matrices de graphiques permettant d'étudier en une seule fois une série de graphiques à deux dimensions.

Remarque : l'outil XLSTAT-3DPlot permet de créer des graphiques beaucoup plus percutants grâce à un grand nombre d'options, avec la possibilité de représenter les données sur un troisième axe.

Dans cette section :

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.

Onglet **Général**:

X : sélectionnez dans ce champ les données à utiliser comme coordonnées pour l'axe des abscisses.

Y : sélectionnez dans ce champ les données à utiliser comme coordonnées pour l'axe des ordonnées.

Z : activez cette option pour sélectionner les données qui conditionneront la taille des points sur les graphiques.

- **Utiliser les bulles** : activez cette option pour utiliser les graphiques avec bulles de MS Excel.

Groupes : activez cette option pour sélectionner les données qui correspondent à l'identifiant du groupe auquel appartient chaque observation. Sur le graphique, la couleur des points dépend du groupe.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des variables : activez cette option si la première ligne des données sélectionnées (X, Y, Z, groups, poids et libellés des observations) contient un libellé.

Libellés des observations : activez cette option si vous voulez utiliser des étiquettes de lignes disponibles. Si l'option « Libellés des variables » est activée, la première cellule de la sélection doit comprendre un en-tête. Si vous n'activez pas cette option, des libellés seront automatiquement créés (Obs1, Obs2, …).

Poids : activez cette option si vous voulez pondérer les observations. Si vous n'activez pas cette option, les poids seront tous considérés comme valant 1. Les poids doivent être impérativement supérieurs ou égaux à 0. Si un en- tête de colonne a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

Onglet **Options**:

Matrice de graphiques : activez cette option pour afficher l'ensemble des combinaisons possibles de variables deux à deux sous la forme d'un tableau à deux entrées, avec en ligne les variables Y et en colonne les X.

- **Histogrammes** : activez cette option pour que, si les variables X et Y sont identiques, XLSTAT affiche les histogrammes des variables sur la diagonale de la matrice de graphiques.
- **Q-Q plots** : activez cette option pour que, si les variables X et Y sont identiques, XLSTAT affiche les Q-Q plots des variables sur la diagonale de la matrice de graphiques.

Effectifs : activez cette option pour afficher les effectifs correspond à chaque point sur les graphiques.

- **Seulement si > 0** : activez cette option pour n'afficher les effectifs que si ils sont strictement supérieurs à zéro.

Ellipses de confiance : activez cette option pour afficher des ellipses de confiance. Choisissez alors l'intervalle de confiance et la distribution à utiliser (Fisher ou Khi^2).

Légende : activez cette option pour afficher la légende du graphique.

Onglet **Couleurs** :

Nombre de groupes : Sélectionnez cette option pour choisir une couleur par groupe. Rentrez alors le nombre de groupes différents de l'analyse.

Groupe : Cliquez sur le groupe pour choisir sa couleur.

Exemple

Un exemple d'utilisation de l'outil Nuages de points est disponible sur le Centre d'aide XLSTAT à l'adresse

http://www.xlstat.com/demo-scatterf.htm

Motion charts

Utilisez cet outil pour explorer l'évolution de plusieurs variables au cours du temps.

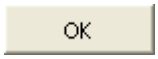
Dans cette section :

Description

Habituellement, lorsque l'on souhaite visualiser l'évolution d'une variable au cours du temps, la valeur de la variable est représentée sur l'axe des ordonnées et la valeur du temps sur l'axe des abscisses. Avec ce type de représentation, on est limité à la visualisation de l'évolution d'une seule variable à la fois. Si l'on souhaite visualiser plusieurs variables, mesurées sur différents individus, plusieurs graphiques sont nécessaires. L'outil Motion Charts de XLSTAT permet de visualiser l'évolution de plusieurs variables (jusqu'à 3), mesurées sur plusieurs individus, au cours du temps sur un seul graphique dynamique.

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.

Onglet **Général**:

X : sélectionnez dans ce champ les données à utiliser comme coordonnées pour l'axe des abscisses.

Y : sélectionnez dans ce champ les données à utiliser comme coordonnées pour l'axe des ordonnées.

Temps : sélectionnez dans ce champ les données de dates ou d'heure, ou une variable numérique quelconque correspondant aux observations de la série temporelle.

Groupes : activez cette option pour sélectionner les données qui correspondent à l'identifiant du groupe auquel appartient chaque observation. Sur le graphique, la couleur des points dépend du groupe.

Taille : activez cette option pour sélectionner les valeurs qui détermineront la taille des points sur le graphique.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des variables : activez cette option si la première ligne des données sélectionnées (X, Y, Temps, Groupes et Taille) contient un libellé.

Interpoler les points manquants : activez cette option pour estimer les coordonnées des points manquants comme la moyenne des coordonnées précédente et suivante.

Lissage : activez cette option pour afficher les positions intermédiaires des points entre deux dates consécutives dans le but de produire une évolution continue entre deux dates.

Légende : activez cette option pour afficher une légende contenant les valeurs minimum et maximum de Taille selon la taille des points.

Onglet **Données manquantes** :

Ne pas accepter les valeurs manquantes : activez cette option pour que XLSTAT empêche la poursuite des calculs si des valeurs manquantes sont détectées

Supprimer les observations : activez cette option pour supprimer les observations comportant des données manquantes.

Remplacer par la moyenne des valeurs précédente et suivante : activez cette option pour estimer les données manquantes par la moyenne de la première valeur précédente non manquante et de la première valeur suivante non manquante.

Boutons associés à un Motion Chart

: la barre de défilement au bas du graphique permet de changer manuellement les valeurs de la variable Temps et ainsi afficher les positions des points à différentes dates.

 : cliquez sur ce bouton pour diminuer automatiquement les valeurs du temps et ainsi voir les points se déplacer lorsque le temps diminue.

 : cliquez sur ce bouton pour arrêter le défilement du temps et ainsi stopper le déplacement des points.

 : cliquez sur ce bouton pour augmenter automatiquement les valeurs du temps et ainsi voir les points se déplacer lorsque le temps augmente.

 : cliquez sur ce bouton pour régler la vitesse de déplacement (de 1, très lent, à 10 très rapide) des points au cours du temps.

Exemple

Un exemple d'utilisation de l'outil Motion charts est disponible sur le Centre d'aide XLSTAT à l'adresse

http://www.xlstat.com/demo-motionf.htm

Bar chart race

Utilisez cet outil pour visualiser l'évolution d'une variable au cours du temps au sein de plusieurs groupes d'observations.

Dans cette section :

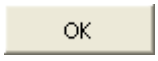
Description

Lorsque l'on souhaite visualiser les valeurs d'une variable sur différentes observations, on peut utiliser un bar chart ou diagramme en bâtons. La hauteur des barres associées à chacune des observations représente la valeur de la variable. Cependant, si l'on dispose de mesures de cette variable à différentes périodes le bar chart classique ne permet pas de représenter l'évolution au cours du temps. Le bar chart race (ou course de bar chart) permet de visualiser l'évolution d'une variable au cours du temps sur un seul graphique dynamique.

Deux options d'affichage des données sont proposées. L'utilisateur peut choisir de représenter les données brutes à chaque intervalle de temps, ou bien d'afficher les données cumulées. L'affichage des données cumulées peut être très utile pour des données de comptage par exemple.

Si par exemple pour un temps t il n'y a pas de données pour un certain groupe, alors la valeur affichée sera la valeur du groupe au temps précédent.

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.

 : cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.

 : cliquez sur ce bouton pour changer la façon dont vous chargez les données dans XLSTAT. Si le bouton possède une icône de souris, XLSTAT vous permet de faire une sélection à la souris des données. Si le bouton possède une icône de liste, XLSTAT vous permet de sélectionner les données à partir d'une liste. Si le bouton possède une icône orange, des boutons complémentaires avec un point d'interrogation  sont affichés pour vous permettre d'importer les données à partir de fichiers.

Onglet **Général**:

Données : sélectionnez dans ce champ la variable que vous souhaitez représenter sur le bar chart.

Temps : sélectionnez dans ce champ les données de dates ou d'heure, ou une variable numérique quelconque correspondant aux observations de la série temporelle.

Groupes : sélectionnez dans ce champ les données qui correspondent à l'identifiant du groupe auquel appartient chaque observation.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des variables : activez cette option si la première ligne des données sélectionnées contient un libellé.

Onglet **Options**:

Position des barres :

- **Position fixe** :
    - **Plus petit en bas** : le groupe avec la plus petite valeur au dernier temps sera affiché tout en bas du graphique.

- **Plus grand en bas** : le groupe avec la plus grande valeur au dernier temps sera affiché tout en bas du graphique.
- **Tri alphabétique** : les groupes sont affichés dans l'ordre alphabétique.

- **Couleurs** :
  - **Couleurs distinctes** : chaque groupe est affiché avec une couleur distincte.
  - **Echelle de couleurs** : une échelle de couleur graduelle est utilisée de sorte que le groupe avec la plus grande valeur finale est affiché en rouge foncé.

- **Données** :
  - **Cumuler** : à chaque intervalle de temps t la valeur affichée correspond à la valeur présente dans vos données au temps t plus la valeur au temps $t-1$.
  - **Données brutes** : à chaque intervalle de temps la valeur affichée correspond à la valeur présente dans vos données.

Onglet **Données manquantes** :

Ne pas accepter les valeurs manquantes : activez cette option pour que XLSTAT empêche la poursuite des calculs si des valeurs manquantes sont détectées

Utiliser la valeur précédente : activez cette option pour remplacer une donnée manquante au temps t par la valeur présente au temps $t-1$.

Boutons associés à un Bar chart race

: la barre de défilement au bas du graphique permet de changer manuellement les valeurs de la variable Temps et ainsi afficher les positions des barres à différentes dates.

<: cliquez sur ce bouton pour diminuer automatiquement les valeurs du temps et ainsi voir les barres évoluer lorsque le temps diminue.

||: cliquez sur ce bouton pour arrêter le défilement du temps et ainsi stopper le déplacement des barres.

>: cliquez sur ce bouton pour augmenter automatiquement les valeurs du temps et ainsi voir les barres évoluer lorsque le temps augmente.

: cliquez sur ce bouton pour régler la vitesse de déplacement (de 1, très lent, à 10 très rapide) des barres au cours du temps.

Exemple

Un exemple d'utilisation de l'outil Bar chart race est disponible sur le Centre d'aide XLSTAT à l'adresse suivante :

http://www.xlstat.com/demo-racf.htm

Graphiques en coordonnées parallèles

Utilisez cet outil pour visualiser des données multidimensionnelles (décrites par p_1 variables quantitatives et p_2 variables qualitatives) sur un même graphique à deux dimensions.

Dans cette section :

Description

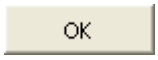
Cette méthode de visualisation est particulièrement utile en analyse de données pour détecter ou pour valider l'existence de groupes homogènes. On peut par exemple utiliser cette méthode à l'issue d'une Classification Ascendante Hiérarchique (CAH).

Si l'on considère que n individus sont décrits par p_1 variables quantitatives et p_2 variables qualitatives, le graphique consiste en $p = p_1 + p_2$ axes verticaux représentant chacun une variable, et n lignes correspondant à chacun des individus. Une ligne croise un axe à la valeur que prend l'individu sur la variable correspondante.

A cause des limitations d'Excel (255 séries) vous ne pouvez pas représenter plus de 255 individus. Si vous sélectionnez plus d'individus, seuls les 255 premiers seront affichés. De plus quand beaucoup d'individus sont présents, le graphique peut vite devenir illisible. C'est pour cela que XLSTAT vous permet de représenter des statistiques descriptives résumant les individus au lieu de représenter tous les individus.

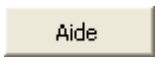
Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.



: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.

 : cliquez sur ce bouton pour effacer les sélections de données.

 : cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.

 : cliquez sur ce bouton pour changer la façon dont vous chargez les données dans XLSTAT. Si le bouton possède une icône de souris, XLSTAT vous permet de faire une sélection à la souris des données. Si le bouton possède une icône de liste, XLSTAT vous permet de sélectionner les données à partir d'une liste. Si le bouton possède une icône orange, des boutons complémentaires avec un point d'interrogation  sont affichés pour vous permettre d'importer les données à partir de fichiers.

Onglet **Général** :

Données quantitatives : activez cette option pour sélectionner les variables quantitatives décrivant les individus. Si des en-têtes de colonnes ont été sélectionnés, veuillez vérifier que l'option « Libellés des variables » est activée.

Données qualitatives : activez cette option pour sélectionner les variables qualitatives décrivant les individus. Si des en-têtes de colonnes ont été sélectionnés, veuillez vérifier que l'option « Libellés des variables » est activée.

Poids : cette option uniquement disponible si vous avez choisi de représenter les lignes de statistiques descriptives permet de pondérer les observations. Si vous n'activez pas cette option, les poids seront tous considérés comme valant 1. Les poids doivent être impérativement supérieurs ou égaux à 0. Si un en-tête de colonne a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

Groupes : activez cette option pour sélectionner les données qui correspondent à l'identifiant du groupe auquel appartient chaque individu. Si des en-têtes de colonnes ont été sélectionnés, veuillez vérifier que l'option « Libellés des variables » est activée. Sur le graphique, la couleur des lignes dépend du groupe.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des variables : activez cette option si la première ligne des données sélectionnées (données quantitatives, qualitatives, poids et groupes et libellés des observations) contient un

libellé.

Libellés des observations : activez cette option si vous voulez utiliser des libellés d'observations disponibles sur une feuille Excel pour l'affichage des résultats. Si l'option « Libellés des variables » est activée, la première cellule de la sélection doit comprendre un en-tête.

Remettre à l'échelle : activez cette option pour que toutes les variables soient représentées sur la même échelle 0%-100% (pour les variables numériques 0 correspond au minimum et 100 au maximum ; pour les variables nominales, les modalités sont régulièrement espacées, et classées en ordre alphabétique).

Onglet **Options** :

Type d'affichage :

- **Une ligne par observation** : activez cette option pour afficher autant de lignes parallèles que possible (le maximum est 250 du fait des limitations d'Excel).
- **Lignes de statistiques descriptives** : activez cette option pour n'afficher que les lignes correspondant aux statistiques suivantes :
- **Minimum et maximum**
- **Médiane**
- **Premier quantile (%)** : entrez la valeur du premier quantile (2.5% par défaut).
- **Deuxième quantile (%)** : entrez la valeur du deuxième quantile (97.5% par défaut).
- **Mode** (pour les variables qualitatives).

Données quantitatives :

- **Données brutes** : activez cette option pour utiliser les données brutes des variables quantitatives. L'échelle sur l'axe Y sera comprise entre le minimum et le maximum de toutes les variables.
- **Remettre à l'échelle** : activez cette option pour que toutes les variables soient représentées sur la même échelle entre 0 (minimum) et 1 (maximum).
- **Axes Y différents** : activez cette option pour que chaque axe Y associé à une variable dispose de sa propre échelle.

Onglet **Données manquantes** :

Ne pas accepter les valeurs manquantes : activez cette option pour que XLSTAT empêche la poursuite des calculs si des valeurs manquantes sont détectées.

Ignorer les données manquantes : si vous choisissez cette option, les données manquantes seront ignorées lors de l'affichage.

Supprimer les observations : activez cette option pour supprimer les observations comportant des données manquantes.

Estimer les données manquantes : activez cette option pour estimer les données manquantes avant le début des calculs.

- **Moyenne ou mode** : activez cette option pour estimer les données manquantes en utilisant la moyenne (variables quantitatives) ou le mode (variables qualitatives) pour les variables correspondantes.
- **Plus proche voisin** : activez cette option pour estimer les données manquantes d'une observation en recherchant le plus proche voisin de l'observation.

Exemple

Un exemple de génération d'un graphique en coordonnées parallèles est disponible sur le Centre d'aide XLSTAT à l'adresse suivante :

http://www.xlstat.com/demo-parf.htm

Bibliographie

Inselberg A. (1985). The Plane with Parallel Coordinates. *The Visual Computer*, **1**, 69-91.

Eickemeyer J. S., Inselberg A., Dimsdale B. (1992). Visualizing p-flats in n-space Using Parallel Coordinates. Technical Report G320-3581, IBM Palo Alto Scientific Center.

Wegman E.J. (1990). Hyperdimensional Data Analysis Using Parallel Coordinates. *J. Amer. Statist. Assoc.*, **85**, 411, 664-675.

Diagrammes ternaires

Utilisez cet outil pour créer des diagrammes ternaires permettant de représenter à l'intérieur d'un triangle des points ayant leurs coordonnées dans un espace à trois dimensions, avec comme contrainte que la somme des coordonnées est constante.

Dans cette section :

Description

Cette méthode de visualisation est particulièrement utile dans les domaines où l'on travaille avec trois éléments dont ont fait varier les proportions. Par exemple en chimie ou en pétrologie.

Cet outil permet de créer très rapidement un diagramme ternaire en représentant des points ainsi que les lignes de projection reliant chaque point à chacun des axes.

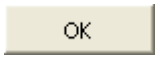
Il existe deux approches pour les graphiques ternaires :

- soit les segments de projection orthogonale des points sur les axes donnent l'information des proportions relatives des 3 éléments.
- soit la projection parallèle à l'axe A permet de lire la coordonnée du point sur l'axe B, l'axe B suivant l'axe A en tournant dans le sens contraire des aiguilles d'une montre.

XLSTAT permet actuellement uniquement la seconde représentation.

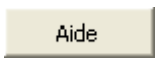
Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.



: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.

 : cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.

Onglet **Général**:

X : sélectionnez les données quantitatives correspondant à la quantité du premier élément.

Y : sélectionnez les données quantitatives correspondant à la quantité du second élément.

Z : activez cette option pour sélectionner les données quantitatives correspondant à la quantité du troisième élément.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des variables : activez cette option si la première ligne des données sélectionnées contient un libellé.

Onglet **Options**:

Constante : veuillez entrer la valeur de la somme constante des coordonnées. La valeur par défaut est 1.

Onglet **Sorties**:

Statistiques descriptives : activez cette option pour afficher les statistiques descriptives pour les variables sélectionnées.

Onglet **Graphiques** :

X1/X2 | Min/Max : vous pouvez modifier ici les min et les max pour les deux premiers axes. Ceux du troisième axe sont déterminés à partir de ceux des deux autres axes. Veillez à ce que la quantité Max-Min soit la même pour les deux axes.

Nombre de segments : veuillez préciser ici en combien de segments les axes doivent-ils être découpés.

Lignes de projection : activez cette option pour afficher les lignes en pointillés rouges reliant les points aux 3 axes indiquant ainsi leurs coordonnées.

Lier le graphique aux données d'entrée : activez cette option lier le graphique aux données d'entrée. Ainsi à chaque changement dans les données d'entrée le graphique répercutera le changement.

Exemple

Un exemple de génération de diagramme ternaire est disponible sur le Centre d'aide XLSTAT à l'adresse suivante :

http://www.xlstat.com/demo-ternaryf.htm

Graphiques 2D pour tableaux croisés

Utilisez cet outil pour créer un graphique bidimensionnel basé sur un tableau de contingence ou un tableau croisé.

Dans cette section :

Description

Cet outil de visualisation permet de générer rapidement un graphique 2D montrant l'importance relative des diverses combinaisons que vous pouvez obtenir lorsque vous créez un tableau de contingence (tableau contenant les comptages d'occurrences de combinaisons de modalités de deux variables qualitatives - par exemple les nombres de personnes répartis par classe d'âge répondant oui ou non à une question) ou plus généralement un tableau croisé (par exemple, un tableau dans lequel est affiché le chiffre d'affaire par pays pour différents produits).

Cet outil peut travailler directement sur les données brutes (pondérées ou non) ou sur un tableau de contingence.

Boîte de dialogue

OK : cliquez sur ce bouton pour lancer les calculs.

Annuler : cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

Aide : cliquez sur ce bouton pour afficher l'aide.

 : cliquez sur ce bouton pour rétablir les options par défaut.

 : cliquez sur ce bouton pour effacer les sélections de données.

 : cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.

Onglet **Général**:

Format des données : choisissez le format des données.

- **Tableau croisé** : activez cette option si vos données sont disponibles sous la forme d'un tableau de contingence ou plus généralement d'un tableau croisé.
- **Variables qualitatives** : activez cette option si vos données se présentent sous la forme de deux variables qualitatives à partir desquelles sera généré un tableau de contingence.

Tableau de contingence : si le format de données choisi est « Tableau de contingence », sélectionnez un tableau croisé, avec les fréquences correspondant aux différentes catégories de deux variables qualitatives. Si les libellés des lignes et des colonnes du tableau ont été sélectionnés, veillez à ce que l'option « Libellés inclus » soit activée.

Variable qualitative(1) : si le format de données choisi est « Variables qualitatives », sélectionnez les données correspondant à la variable qualitative qui sera en ligne dans le tableau de contingence et qui correspondra aux ordonnées sur le graphique. Si le libellé de la variable a été sélectionné, veillez à ce que l'option « Libellés des variables » soit bien activée.

Variable qualitative(2) : si le format de données choisi est « Variables qualitatives », sélectionnez les données correspondant à la variable qualitative qui sera en colonne dans le tableau de contingence créé et qui correspondra aux abscisses sur le graphique. Si le libellé de la variable a été sélectionné, veillez à ce que l'option « Libellés des variables » soit bien activée.

Z : si le format de données choisi est « Variables qualitatives »", activez cette option puis sélectionnez les valeurs qui pondèreront les observations avec un impact sur la taille des points sur le graphique. Vous pouvez sélectionnez plusieurs colonnes. Si vous voulez que les points soient sur la même échelle pour les différentes dimensions, activez l'option **même échelle**.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés inclus : activez cette option si les libellés des lignes et des colonnes du tableau de contingence ont été sélectionnés.

Libellés des variables : activez cette option si la première ligne des données sélectionnées contient un libellé.

Afficher un titre : Activez cette option pour afficher un titre sur le graphique.

Onglet **Options**:

Forme : Choisissez la forme que vous voulez utiliser.

- **Cercle**

- **Carré**
- **Bulles**

Afficher les valeurs : Activez cette option si vous souhaitez afficher les valeurs sur le graphique.

Taille : choisissez quel paramètre des points est lié à la dimension Z ou aux valeurs du tableau de contingence, entre la surface, la largeur ou la largeur au carré. Si vous choisissez l'option surface la différence entre les points les plus gros et les plus petits sera atténuée.

Echelle(%) : choisissez le facture de mise à l'échelle des points. La valeur par défaut est 100. Le même facteur est appliqué à tous les points.

Axe horizontal en bas : activez cette option pour utiliser les graphiques avec bulles de MS Excel.

Quadrillage : activez cette option pour afficher le quadrillage sur le graphique.

Exemple

Un exemple de création d'un graphique à partir d'un tableau de contingence est proposé sur le Centre d'aide XLSTAT :

http://www.xlstat.com/demo-2dcontf.htm

Barres d'erreur

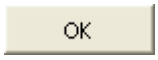
Utilisez cet outil pour créer facilement des graphiques Excel avec des barres d'erreur pouvant être différentes pour chaque point.

Dans cette section :

Description

Cet outil vient palier un défaut d'Excel : s'il est possible d'ajouter des barres d'erreur sur différent types de graphiques, cette opération s'avère fastidieuse si les bornes ne sont pas les mêmes pour tous les points. Avec cet outil vous pouvez créer un graphique avec des barres d'erreur en une seule fois.

Boîte de dialogue



: cliquez sur ce bouton pour créer les graphiques.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer aucune opération.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes (mode colonnes). Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes (mode lignes).

Onglet **Général**:

X : sélectionnez dans ce champ les données à utiliser comme coordonnées pour l'axe des abscisses. Si vous sélectionnez plusieurs colonnes (mode colonnes) ou plusieurs lignes (mode lignes) vous devez ensuite sélectionner le même nombre de colonnes (ou de lignes) pour les Y, les bornes inférieures et supérieures. En revanche, si vous ne sélectionnez qu'une seule

colonne (ou ligne) vous pouvez ensuite sélectionner une ou plusieurs colonnes (ou lignes) pour les Y, les bornes inférieures et supérieures.

Y : sélectionnez dans ce champ les données à utiliser comme coordonnées pour l'axe des ordonnées. Voir ci-dessus les contraintes quant au nombre de colonne (ou lignes) à sélectionner.

Borne inférieure : activez cette option si vous souhaitez ajouter des bornes inférieures sur le graphique. Sélectionnez alors dans ce champ les données à utiliser comme bornes inférieures. Le nombre de colonnes (mode colonnes) ou lignes (mode lignes) à sélectionner doit être égal à celui des Y.

Borne supérieure : activez cette option si vous souhaitez ajouter des bornes supérieures sur le graphique. Sélectionnez alors dans ce champ les données à utiliser comme bornes supérieures. Le nombre de colonnes (mode colonnes) ou lignes (mode lignes) à sélectionner doit être égal à celui des Y.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des variables : activez cette option si la première ligne des données sélectionnées (données quantitatives, qualitatives, poids et groupes et libellés des observations) contient un libellé.

Onglet **Graphiques** :

Type de graphique : choisissez le type de graphique à afficher :

- Diagramme en bâtons.
- Courbe.
- Nuage de points.

Exemple

Un exemple de création d'un graphique avec des barres d'erreur est disponible sur le Centre d'aide XLSTAT :

http://www.xlstat.com/demo-errf.htm

Nuage de mots

Cette méthode permet de créer une représentation visuelle de données textuelles. Elle est généralement utilisée pour représenter des mots clés en fonction de leur fréquence dans un ou plusieurs documents.

Les termes sont généralement des mots seuls, et l'importance de chacun est représentée en faisant varier la taille de la police et / ou la couleur.

Ce graphique est utile pour identifier rapidement les termes les plus importants d'un document, ainsi que l'importance relative des termes.

Dans cette section :

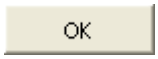
Description

Le nuage de mots est un outil de visualisation basé sur la fréquence. Dans sa forme la plus simple, une seule dimension de l'information est représentée : la taille de la police est proportionnelle à la fréquence des mots, ce qui signifie que plus un mot est grand dans le nuage de mots, plus le mot apparait fréquemment dans le document.

La fonctionnalité « nuage de mots » nécessite comme données d'entrée un vecteur de termes auquel s'ajoute une matrice de fréquence des termes (avec une colonne par document), et affiche en sortie un nuage de mots par document.

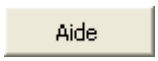
Une échelle de couleurs personnalisée peut être spécifiée afin de contrôler la couleur des mots du moins, tout en gardant celle-ci dépendante de la fréquence relative des termes.

Boîte de dialogue



: cliquez sur ce bouton pour créer les graphiques.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer aucune opération.



: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.

 : cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes (mode colonnes). Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes (mode lignes).

Onglet **Général** :

Matrice fréquence terme : Sélectionnez ici la matrice de fréquence des termes. Si un en-tête a été sélectionné, vérifiez que l'option "Libellés des colonnes/lignes" a été activée.

Etiquettes de terme : Sélectionnez le vecteur des étiquettes de terme dans ce champ. Si un en-tête fait partie de la sélection, vérifiez que l'option "Libellés des colonnes/lignes" a été activée.

Colorer par : * **Fréquence** : Choisissez cette option pour que la couleur des mots dépende de leur fréquence. * **Valeurs** : Choisissez cette option pour que la couleur des mots dépende d'une valeur que vous fournissez. Dans ce cas, vous devez sélectionner les données correspondantes. Si l'option "libellés des colonnes/lignes" est cochée, assurez-vous de sélectionner également un en-tête pour ce champ.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des colonnes/lignes : Activez cette option si la première ligne (ou colonne si les données sont transposées) des données sélectionnées (matrice fréquence terme, étiquettes de terme, valeurs) contient un en-tête.

Onglet **Options** :

Mots max. : Activez cette option pour fixer le nombre maximum de mots à afficher. Les termes les moins fréquents ne sont pas affichés.

Position aléatoire : Activez cette option pour afficher les mots dans un ordre aléatoire. Si l'option est désactivée, ils seront affichés par ordre de fréquence décroissant, ainsi le mot le plus fréquent se trouvera au centre du graphique, entouré des mots dont la fréquence est immédiatement inférieure, etc.…

Période de rot. : Activez cette option pour définir la période de rotation de 90° entre chaque mot affiché. Par défaut la période de rotation est de 4 mots, c'est-à-dire qu'un mot sur quatre subira une rotation de 90° dans la séquence d'affichage du nuage.

Echelle de couleur : Sélectionnez le type d'échelle de couleurs.

- **Prédéfinie** : Activez cette option pour choisir une échelle de couleurs prédéfinie. Activez l'option **Échelle log** si vous souhaitez accentuer la différence de couleur entre les termes de faible fréquence.
- **Echelle de couleurs aléatoire** : Activez cette option pour que les couleurs soient affectées aux mots de façon aléatoire
- **Echelle de couleurs personnalisée** : Activez cette option pour sélectionner les cellules correspondant à l'échelle de couleurs (en fréquence décroissante)

Résultats

Pour chaque document sélectionné (chaque colonne de la matrice fréquence terme), un nuage de mots est affiché. Le nuage de mots est un graphique Excel. Vous pouvez donc modifier les couleurs, positions, police de caractère, taille comme vous le voulez.

Exemple

Un exemple de création d'un nuage de mots est disponible sur :

http://www.xlstat.com/demo-wdcf.htm

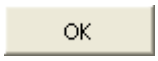
Graphiques en radar

Utilisez cet outil pour créer des graphiques en radar (ou graphiques en toile d'araignée, ou graphiques en étoile). Ce type de graphique permet d'évaluer différents choix en fonction de plusieurs variables. Ils peuvent être utiles par exemple pour comparer des températures à plusieurs endroits au cours d'une année, pour évaluer la qualité, pour comparer des notes à une moyenne, etc. XLSTAT permet d'afficher les libellés autour du graphique soit sous forme de texte soit sous forme d'images.

Dans cette section :

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.

Onglet **Général** :

Variables ou **Images** : XLSTAT permet de sélectionner soit des noms de variables sous deux formes : sous forme de texte ou sous forme d'images. Sélectionnez dans ce champ les libellés ou les images, en fonction de l'option choisie, qui apparaitront autour du graphique.

Données : sélectionnez dans ce champ les données numériques pour chaque variable.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des variables : activez cette option si la première ligne des données sélectionnées contient un libellé.

Onglet **Options** :

Type de graphique : XLSTAT permet d'afficher les graphiques en radar sous différentes formes : * **Classique** : activez cette option pour afficher un graphique en radar classique.

- **Avec des points** : activez cette option pour afficher un graphique en radar avec des points.
- **Plein** : activez cette option pour afficher un graphique ou l'intérieur du graphique est plein, avec une couleur par catégorie.
- **Graphique polaire** : activez cette option pour afficher un graphique polaire.

Grouper les variables : activez cette option pour regrouper les variables sur un même graphique.

Etiquettes : activez cette option pour afficher les étiquettes sur les graphiques.

Valeurs d'axe : activez cette option pour afficher les valeurs d'axe sur le graphique.

Exemple

Un exemple d'utilisation de l'outil Graphiques en radar est disponible sur le Centre d'aide XLSTAT à l'adresse

http://www.xlstat.com/radar_fr.htm

Diagrammes en tornade

Utiliser cet outil pour créer un graphique en tornade ou un histogramme dos à dos à partir d'un échantillon de données quantitatives continues ou discrètes.

Dans cette section :

Description

Diagrammes en tornade

Le diagramme en tornade est un outil de visualisation, similaire à un graphique en barres, qui permet de comparer l'importance relative de deux variables. Les catégories sont en général ordonnées de sorte que la plus grande barre apparaisse en haut du graphique, la deuxième plus grande en deuxième et ainsi de suite, mais XLSTAT permet de ne pas ordonner les catégories si vous le souhaitez. Au final, le graphique ressemble à une tornade, d'où son nom.

Histogrammes

L'histogramme est l'un des outils de visualisation les plus utilisés car il permet d'avoir très rapidement une idée de la distribution d'un échantillon de données quantitatives continues ou discrètes.

Dans le cas où l'on a deux variables ou deux groupes, il est possible d'afficher un histogramme dos à dos. Ceci peut notamment permettre d'identifier les différences dans la tendance centrale, la variabilité, l'asymétrie et la forme, ainsi que les valeurs aberrantes ou les lacunes dans les données, ce qui conduira à des informations plus précises et plus précieuses.

Définition des intervalles

L'un des enjeux pour la création d'un histogramme est la définition des intervalles, car pour un jeu de données déterminé, l'allure de l'histogramme en dépend entièrement. Entre les deux extrêmes de l'intervalle unique comprenant toutes les données et donnant une seule barre, et de l'histogramme où il y a un intervalle par donnée, il existe autant d'histogrammes possibles que de partitions des données.

Afin d'obtenir un résultat visuellement et/ou opérationnellement satisfaisant, la définition des intervalles peut nécessiter plusieurs aller-retours.

La méthode la plus classique consiste à utiliser des intervalles de même amplitude, la valeur du premier intervalle étant déterminée par la valeur minimale ou une valeur légèrement inférieure.

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.

OK : cliquez sur ce bouton pour lancer les calculs.

Annuler : cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

Aide : cliquez sur ce bouton pour afficher l'aide.

 : cliquez sur ce bouton pour rétablir les options par défaut.

 : cliquez sur ce bouton pour effacer les sélections de données.

 : cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.

 : cliquez sur ce bouton pour changer la façon dont vous chargez les données dans XLSTAT. Si le bouton possède une icône de souris, XLSTAT vous permet de faire une sélection à la souris des données. Si le bouton possède une icône de liste, XLSTAT vous permet de sélectionner les données à partir d'une liste. Si le bouton possède une icône orange, des boutons complémentaires avec un point d'interrogation  sont affichés pour vous permettre d'importer les données à partir de fichiers.

Onglet **Général** :

Format :

Histogrammes : choisissez cette option pour que XLSTAT affiche un histogramme dos à dos.

Diagrammes en tornade : choisissez cette option pour que XLSTAT affiche un diagramme en tornade.

Données : sélectionnez deux colonnes de données quantitatives ou bien une colonne de données quantitatives et une colonne avec deux groupes. Si des en-têtes ont été sélectionnés, veuillez vérifier que l'option « Libellés des colonnes » est activée.

Dans le cas où l'on choisit d'afficher un histogramme dos à dos :

Type de données :

Continues : choisissez cette option pour que XLSTAT considère que vos données sont continues.

Discrètes : choisissez cette option pour que XLSTAT considère que vos données sont discrètes.

Dans le cas où l'on choisit d'afficher un diagramme en tornade :

Etiquettes : sélectionnez cette option si vous souhaitez entrer une colonne de labels.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des colonnes : activez cette option si la première ligne des données sélectionnées contient un libellé.

Groupes : activez cette option puis sélectionnez une colonne (mode colonnes) ou une ligne (mode lignes) contenant les descripteurs des données. Si un en-tête a été sélectionné, veuillez vérifier que l'option « Libellés des échantillons » est activée.

Onglet **Options** :

Titre du graphique : entrez le titre du graphique à créer.

Dans le cas où l'on choisit d'afficher un diagramme en tornade :

Barres : choisissez si vous voulez créer un diagramme avec des barres **verticales** ou **horizontales**.

Trier : choisissez cette option si vous souhaitez trier les données du diagramme en tornade.

Onglet **Sorties** :

Statistiques descriptives : activez cette option pour afficher les statistiques descriptives des échantillons.

Résultats

Statistiques simples : dans ce tableau sont affichées pour tous les échantillons les statistiques descriptives suivantes : le nombre d'observations, le nombre de données manquantes, le nombre de données non manquantes, la moyenne, et l'écart-type (non biaisé).

Diagrammes en tornade : les diagrammes en tornade sont affichés. Si vous le souhaitez, vous pouvez modifier la couleur des lignes, les échelles, et les titres comme avec n'importe quel graphique Excel.

Histogrammes : les histogrammes sont affichés. Si vous le souhaitez, vous pouvez modifier la couleur des lignes, les échelles, et les titres comme avec n'importe quel graphique Excel.

Statistiques descriptives pour les intervalles : dans ce tableau sont affichés pour chaque intervalle sa borne inférieure, sa borne supérieure, le nombre de valeurs de l'échantillon étant comprises dans l'intervalle (effectif), la fréquence (l'effectif divisé par l'effectif total de l'échantillon), et la densité (le rapport de la fréquence sur la taille de l'intervalle).

Exemple

Un exemple de génération de diagrammes en tornade est disponible sur le Centre d'aide XLSTAT à l'adresse suivante :

http://www.xlstat.com/demo-tornado.htm

Diagrammes en bâtons

Utilisez cet outil pour afficher rapidement des diagrammes en bâtons qui donne la possibilité d'utiliser soit des étiquettes classiques, soit des images comme étiquettes et / ou comme arrière-plan des barres. Si vos données correspondent à des pays, vous pouvez utiliser automatiquement la bibliothèque de drapeaux de XLSTAT. Vous pouvez également utiliser vos propres images.

In this section:

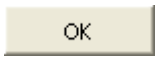
Description

Cet outil a été développé pour vous aider à créer rapidement des graphiques plus explicites.

Si vous avez des demandes pour plus d'options, faites-le nous savoir, nous serons heureux de vous aider à gagner plus de temps.

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.

Onglet **Général** :

Etiquettes : Sélectionnez les données qui décrivent les valeurs et qui peuvent être utilisées comme étiquettes sur les graphiques.

- **Pays (Noms)** : choisissez cette option si les *étiquettes* correspondent aux noms de pays.
- **Pays (Codes)** : choisissez cette option si les *étiquettes* correspondent aux codes (ISO 3166, deux lettres) de pays.
- **Autre** : choisissez cette option si les *étiquettes* ne correspondent pas à des pays. Cette option nécessite que vous disposiez d'un ensemble d'images que vous pouvez sélectionner dans le champ *Images*.

Valeurs : sélectionnez les données qui sont à utiliser pour les barres du diagramme.

Images : sélectionnez les cellules qui pour chaque étiquette contienne l'image à utiliser.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des lignes/colonnes : activez cette option si la première ligne/colonne des données sélectionnées comprend un libellé.

Taille des graphiques : * **Largeur** : entrez la valeur en points de la largeur du graphiques. * **Hauteur** : entrez la valeur en points de la hauteur du graphique.

Onglet **Options** :

Titre du graphique : entrez le titre du graphique à créer.

Barres : choisissez si vous voulez créer un diagramme avec des barres **verticales** ou **horizontales**.

Etiquettes : choisissez si vous souhaitez afficher les libellés ou les codes (l'option codes n'est disponible que si vous avez sélectionné des pays comme libellés) sur les graphiques. Vous pouvez choisir de ne rien à afficher. Dans ce cas, seules les images seront affichées.

Position des images : vous pouvez afficher les images à côté de l'axe ou dans les barres elle-mêmes.

- **A côté de l'axe** : cochez cette option si vous souhaitez afficher les images à côté de l'axe. Si vos libellés correspondent à des pays, vous pouvez choisir d'afficher les drapeaux dans des cercles, des carrés ou des rectangles.
- **Images dans les barres**: Cochez cette option si vous souhaitez afficher les images dans les barres elles-mêmes. Vous pouvez afficher des images étirées ou en mosaïque.

Résultats

XLSTAT affiche un diagramme en bâtons pour chaque série sélectionnée dans le champ *Valeurs*.

Exemple

Un exemple montrant comment afficher des diagrammes en bâtons utilisant des images est disponible sur le Centre d'aide XLSTAT à l'adresse suivante :

http://www.xlstat.com/demo-barchartimagesf.htm

Diagrammes en bâtons tronqués

Utilisez cet outil pour créer des diagrammes en bâtons tronqués sur lesquelles l'échelle des valeurs est comprimée sur une portion définie.

In this section:

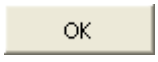
Description

Cet outil de visualisation permet d'afficher efficacement des valeurs qui sont réparties sur deux extrêmes d'une même échelle, sans avoir à utiliser une échelle logarithmique. Bien que ce soit très pratique, il est très fastidieux de créer un tel graphique dans Excel. De plus, XLSTAT permet d'utiliser la transparence pour montrer en partie les informations qui sont cachées dans la partie retirée ou comprimée de l'échelle.

Cet outil permet de travailler directement sur des données quantitatives (typiquement des comptages ou des fréquences), ou sur des données qualitatives brutes (pondérées ou non) qui sont alors transformées en un tableau de comptages avant l'affichage du diagramme en bâtons.

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.

Onglet **Général** :

Format des données: choisissez le format des données.

- **Quantitatives** : activez cette option si vos données correspondent à une simple liste de valeurs numériques, un tableau de contingence ou un tableau croisé. Si l'option "Libellés inclus" est activée, vous devez sélectionner à la fois les libellés d'en-tête (colonnes) et de catégorie (lignes) avec les données.

- **Qualitatives** : activez cette option si vos données sont disponibles sur la forme de variables qualitatives qui doivent être transformées en fréquences avant la création du graphique.

Données quantitatives : si vos données sont quantitatives, sélectionnez les données qui correspondent aux données quantitatives qui doivent être affichées sur le diagramme en bâtons.

- **Un graphique** : activez cette option si vous souhaitez que les résultats correspondant à chacune des colonnes de données quantitatives soient affichés sur un seul graphique.

Données qualitatives : si vos données sont qualitatives, sélectionnez les données qui correspondent aux données qualitatives qui seront utilisées pour construire un tableau de fréquences, qui sera ensuite affiché sur le diagramme en bâtons.

- **Trier les modalités alphabétiquement** : activez cette option to sort alphabetically the categories of the qualitative variables. If this option is unchecked, the order of appearance is respected.

Sous-échantillons : activez cette option pour sélectionner des données correspondant aux noms ou identifiants des sous-échantillons pour chacune des observations.

- **Trier les modalités alphabétiquement** : activez cette option pour trier par ordre alphabétique les modalités des données du sous-échantillon. Si cette option est décochée, l'ordre d'apparition est respecté.

- **Libellés Variable-Modalité**: activez cette option pour utiliser des étiquettes variable-modalité lors de l'affichage des sorties. Les étiquettes Variable-modalités incluent le nom de la variable en tant que préfixe et le nom de la modalité en tant que suffixe.

- **Un graphique** : activez cette option si vous souhaitez que les résultats correspondant à chacun des sous-échantillons soient affichés sur un seul graphique.

Poids : activez cette option si vous voulez pondérer les observations. Si vous n'activez pas cette option, les poids seront tous considérés comme valant 1. Les poids doivent être impérativement supérieurs ou égaux à 0. Si un en- tête de colonne a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés inclus: activez cette option si les libellés des lignes et des colonnes ont été sélectionnés avec les données.

Variable labels: activez cette option si la première ligne des sélections comprend un libellé.

Afficher le titre: activez cette option pour afficher le titre de votre choix sur le graphique.

Tronquer: activez pour comprimer l'axe vertical du graphique (ou l'axe horizontal si l'option "horizontal" est sélectionnée). Sélectionnez alors les points de début et de fin sur l'échelle que vous souhaitez comprimer ou masquer.

Transparence(%): définissez le niveau de transparence de la partie comprimée du diagramme. Choisissez 0 pour complètement cacher l'intervalle ou 100 pour le rendre visible.

Onglet **Sorties** :

Statistiques descriptives : activez cette option pour afficher les statistiques descriptive pour les variables qualitatives sélectionnées.

Exemple

Un tutoriel expliquant comment générer un diagramme en bâtons tronqué est disponible sur :

http://www.xlstat.com/demo-bartruncf.htm

Tracer une fonction

Utilisez cet outil pour tracer une courbe sur un graphique existant ou sur un nouveau graphique.

Dans cette section :

Description

Cet outil permet de tracer une fonction de type $y = f(x)$ sur un graphique existant ou sur un nouveau graphique. La syntaxe de la fonction entrée doit respecter les conventions imposées par Excel pour les fonctions utilisées dans les feuilles de calcul. Par ailleurs, les abscisses doivent être déclarées par X1.

Exemples :

Fonction	Syntaxe XLSTAT
$Y = x^2$	X1²
$Y = ln(x)$	LN(X1)
$Y = e(x)$	EXP(X1)
$Y =\| x \|$	ABS(X1)
$Y = x$ si $x < 0$, $Y = 2x$ si $x = 0$	SI(X1<0; X1; 2*X1)

Par ailleurs, vous pouvez aussi utiliser des fonctions de feuille de calcul XLSTAT. Par exemple, pour tracer la fonction de répartition de la loi normale standard, entrez XLSTAT_CDFNormal(X1).

Boîte de dialogue

OK : cliquez sur ce bouton pour créer les graphiques.

Annuler : cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer aucune opération.

Aide : cliquez sur ce bouton pour afficher l'aide.

 : cliquez sur ce bouton pour rétablir les options par défaut.

 : cliquez sur ce bouton pour effacer les sélections de données.

Onglet **Général**:

Fonction Y = : entrez la fonction que vous voulez tracer, en respectant la syntaxe définie dans la section Description.

Minimum : indiquez la valeur minimale pour laquelle la fonction doit être évaluée et tracée.

Maximum : indiquez la valeur maximale pour laquelle la fonction doit être évaluée et tracée.

Nombre de points : entrez le nombre de points auxquels la fonction doit être évaluée entre les valeurs minimale et maximale saisies. Cette option vous permet d'ajuster la qualité du graphique. Pour une fonction ayant beaucoup de points d'inflexion, trop peu de points risque de donner un graphique de mauvaise qualité. Trop de points risque aussi de détériorer la qualité de l'affichage.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Graphique actif : activez cette option pour ajouter la fonction sur le graphique actuellement sélectionné.

Exemple

Un exemple de création d'un graphique avec des barres d'erreur est disponible sur le Centre d'aide XLSTAT :

http://www.xlstat.com/demo-funf.htm

AxesZoomer

Utilisez cet outil pour modifier les valeurs minimales et maximales des axes des abscisses et des ordonnées d'une graphique.

Dans cette section :

Boîte de dialogue

Important : avant de lancer cet outil, vous devez sélectionner un graphique de type nuage de points ou courbe.

Appliquer : cliquez sur ce bouton pour appliquer les changements au graphique.

Terminer : cliquez sur ce bouton pour fermer la boîte de dialogue.

Aide : cliquez sur ce bouton pour afficher l'aide.

Min X : entrez la valeur minimale de l'axe des abscisses.

Max X : entrez la valeur maximale de l'axe des abscisses.

Min Y : entrez la valeur minimale de l'axe des ordonnées.

Max Y : entrez la valeur maximale de l'axe des ordonnées.

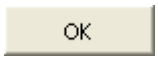
EasyLabels

Utilisez cet outil pour ajouter des étiquettes, éventuellement formatées, à une série de données sur un graphique.

Dans cette section :

Boîte de dialogue

Important : avant de lancer cet outil, vous devez sélectionner un graphique de type nuage de points ou courbe ou une série de points sur un graphique.



: cliquez sur ce bouton pour ajouter les étiquettes.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer de modification.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les étiquettes sont dans une colonne. Si la flèche est vers la droite, XLSTAT considère que les étiquettes sont dans une ligne.

Etiquettes : sélectionnez les étiquettes à ajouter à la série de données sélectionnée sur le graphique.

En-tête dans la première cellule : activez cette option si la première cellule des étiquettes sélectionnées correspond à un en-tête et non à une étiquette.

Utiliser les propriétés du texte : activez cette option si vous souhaitez que le format appliqué au texte contenu dans les cellules contenant les étiquettes soit aussi appliqué au texte des étiquettes sur le graphique :

- **Police** : activez cette option pour utiliser la même police de caractères.
- **Taille** : activez cette option pour utiliser la même taille de police de caractères.

- **Style** : activez cette option pour utiliser le même style de police de caractères (normal, gras, italique).
- **Couleur** : activez cette option pour utiliser la même couleur de police de caractères.

Utiliser les propriétés des cellules : activez cette option si vous souhaitez que le format appliqué aux cellules contenant les étiquettes soit aussi appliqué aux étiquettes sur le graphique :

- **Bordure** : activez cette option pour utiliser la même bordure.
- **Motifs** : activez cette option pour utiliser le même motif.

Utiliser les propriétés des points : activez cette option si vous souhaitez que la couleur des étiquettes soit identique à celle des points :

- **Couleur de l'intérieur** : activez cette option pour utiliser la couleur de l'intérieur des points.
- **Couleur de la bordure** : activez cette option pour utiliser la couleur de la bordure des points.

Repositionnement des étiquettes

Utilisez cet outil pour modifier la position des étiquettes des observations sur un graphique.

Dans cette section :

Boîte de dialogue

OK : cliquez sur ce bouton pour repositionner les étiquettes.

Annuler : cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer de modification.

Aide : cliquez sur ce bouton pour afficher l'aide.

 : cliquez sur ce bouton pour rétablir les options par défaut.

Coins : activez cette option pour placer les étiquettes dans la direction du coin du quadrant dans lequel se trouve le point.

Distance au point :

- **Automatique** : activez cette option pour que XLSTAT détermine automatiquement la distance au point la plus appropriée.
- **Définie par l'utilisateur** : activez cette option pour entrer la valeur (en pixels) de la distance entre l'étiquette et le point.

En haut : activez cette option pour placer les étiquettes au-dessus du point.

A droite : activez cette option pour placer les étiquettes à droite du point.

En bas : activez cette option pour placer les étiquettes au-dessous du point.

A gauche : activez cette option pour placer les étiquettes à gauche du point.

Appliquer uniquement à la série sélectionnée : activez cette option ne modifier l'emplacement des étiquettes que pour la série sélectionnée.

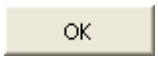
EasyPoints

Utilisez cet outil pour modifier la taille, la couleur ou la forme des points affichés sur un graphique Excel.

Dans cette section :

Boîte de dialogue

Important : avant de lancer cet outil, vous devez sélectionner un graphique de type nuage de points ou courbe ou une série de points sur un graphique.



: cliquez sur ce bouton pour ajouter les étiquettes.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer de modification.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les étiquettes sont dans une colonne. Si la flèche est vers la droite, XLSTAT considère que les étiquettes sont dans une ligne.

Taille : activez cette option et sélectionnez les cellules indiquant la taille à appliquer aux points.

En-tête dans la première cellule : activez cette option si la première cellule des étiquettes sélectionnées correspond à un en-tête et non à une cellule utilisée pour déterminer la taille, la couleur, ou le motif.

Remettre à l'échelle : choisissez l'intervalle de tailles de points à utiliser pour représenter les points. Le minimum est compris entre 2 et 71, et le maximum entre 3 et 72.

Motifs et/ou couleur s : activez cette option et sélectionnez les cellules indiquant quels motifs et/ou couleurs doivent être utilisés pour les points. En ce qui concerne les motifs, 1 correspond

à un carré, 2 à un losange, 3 à un triangle, 4 à un x, 5 à une étoile, 6 à un point, 7 à un -, 8 à un + et 9 à un rond. La couleur de l'intérieur d'un point est définie par la couleur de la cellule et celle du pourtour par la couleur de la bordure des cellules (Remarque : la couleur par défaut des cellules est « sans couleur », donc pour obtenir un point blanc, il faut mettre la couleur du fond à blanc). La taille des points est déterminée à partir des valeurs contenues dans les cellules

Changer les motifs : activez cette option si vous souhaitez que les motifs dépendent des valeurs sélectionnées dans le champ « Motifs et/ou couleurs ».

Utiliser les propriétés des cellules : activez cette option si vous souhaitez que la couleur des cellules soit aussi appliquée aux points :

- **Bordure** : activez cette option pour utiliser la couleur des bordures.
- **Fond** : activez cette option pour utiliser la couleur de fond des cellules.

Exemple

Un exemple d'utilisation de l'outil EasyPoints est disponible sur le Centre d'aide XLSTAT à l'adresse suivante :

http://www.xlstat.com/demo-easypf.htm

Couleurs, épaisseurs et tailles

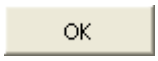
Utilisez cet outil pour modifier la couleur ou l'épaisseur des lignes ou la taille des points de plusieurs séries, en un clic.

Dans cette section :

Description

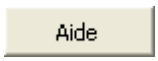
Cet outil vous permet de modifier la couleur et/ou l'épaisseur des lignes et la taille des points de plusieurs séries, en une seule opération.

Boîte de dialogue



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.



: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.

Couleurs : sélectionnez les cellules contenant les couleurs à appliquer aux lignes ou aux points. Si votre graphique comporte S séries, vous devez sélectionner S cellules. La couleur de leur fond sera appliquée aux séries. Si vous voulez colorer chacun des P points des S séries, vous devez sélectionner un tableau avec S lignes et P colonnes.

Epaisseur des lignes : sélectionnez les épaisseurs (en unité de points Excel) des séries dans des cellules. Si l'épaisseur est unique, vous pouvez l'entrer dans le champ de sélection.

Taille des points : sélectionnez les tailles (en unité de points Excel) des points dans des cellules. Si la taille est unique, vous pouvez l'entrer dans le champ de sélection.

En-tête dans la première cellule : activez cette option si la première cellule des données sélectionnées correspond à un en-tête.

Exemple

Un exemple d'utilisation de cette fonction est disponible à l'adresse suivante :

http://www.xlstat.com/demo-cosf.htm

Graphiques orthonormés

Utilisez cet outil pour ajuster le minimum et le maximum de l'axe des abscisses et de l'axe des ordonnées d'un graphique de telle sorte que le graphique soit orthonormé. Cet outil sera particulièrement utile si vous avez agrandi un graphique orthonormé produit par XLSTAT (par exemple après une ACP), et si vous voulez vous assurer que le graphique est toujours orthonormé.

Remarque : un graphique orthonormé est tel qu'une unité en abscisse est visuellement identique à une unité en ordonnée. Les graphiques orthonormés permettent d'éviter des erreurs d'interprétation dues à des effets de dilatation ou d'écrasement.

Dans cette section :

Boîte de dialogue

Appliquer : cliquez sur ce bouton pour appliquer la transformation au graphique.

Annuler : cliquez sur ce bouton pour annuler la transformation du graphique.

Terminer : cliquez sur ce bouton pour fermer la boîte de dialogue.

Aide : cliquez sur ce bouton pour afficher l'aide.

Redimensionner un graphique

Utilisez cet outil pour redimensionner un graphique, ou la zone du graphique délimitée par les axes (zone de traçage).

Dans cette section :

Boîte de dialogue

Appliquer : cliquez sur ce bouton pour redimensionner le graphique.

Terminer : cliquez sur ce bouton pour fermer la boîte de dialogue.

Aide : cliquez sur ce bouton pour afficher l'aide.

Choisissez le type de zone à redimensionner :

- **Graphique** : activez cette option pour redimensionner tout le graphique.
- **Zone de traçage** : activez cette option pour redimensionner uniquement la zone de traçage à l'intérieur du graphique.

Taille actuelle : la largeur et la hauteur affichées ici sont celles du graphique ou de la zone de traçage tels qu'ils sont avant le redimensionnement.

Nouvelle taille : entrez la nouvelle largeur et la nouvelle hauteur du graphique, soit en pourcentage de la taille actuelle, soit en pixels.

Verrouiller les proportions : activez cette option si vous voulez que les proportions initiales du graphique soient respectées.

Transformations de graphiques

Utilisez cet outil pour appliquer une ou plusieurs transformations aux points contenus dans un graphique.

Dans cette section :

Boîte de dialogue

Important : avant de lancer cet outil, vous devez sélectionner un graphique de type nuage de points ou courbe.

OK : cliquez sur ce bouton pour transformer le graphique.

Annuler : cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer de transformation.

Aide : cliquez sur ce bouton pour afficher l'aide.

 : cliquez sur ce bouton pour rétablir les options par défaut.

 : cliquez sur ce bouton pour effacer les sélections de données.

Symétrie :

- **Axe horizontal** : activez cette option pour appliquer une symétrie par rapport à l'axe des abscisses.
- **Axe vertical** : activez cette option pour appliquer une symétrie par rapport à l'axe des ordonnées.

Remarque : si vous sélectionnez les deux options précédentes, la symétrie appliquée sera une **symétrie centrale**.

Translation :

- **Horizontale** : activez cette option pour entrer le nombre d'unités pour la translation horizontale.
- **Verticale** : activez cette option pour entrer le nombre d'unités pour la translation verticale.

Rotation :

- **Angle (°)** : entrez l'angle en degrés pour la rotation à appliquer.
- **Droite** : si cette option est activée la rotation est appliquée dans le sens des aiguilles d'une montre.
- **Gauche** : si cette option est activée la rotation est appliquée dans le sens inverse des aiguilles d'une montre.

Homothétie :

- **Facteur** : entrez le facteur d'homothétie à appliquer aux données.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Afficher les nouvelles coordonnées : activez cette option pour afficher les coordonnées une fois toutes les transformations appliquées.

Mettre à jour le min et le max : activez cette option pour que XLSTAT adapte automatiquement le minimum et le maximum de l'axe des abscisses et de l'axe des ordonnées, une fois les transformations effectuées, de telle sorte que tous les points soient visibles.

Graphique orthonormé : activez cette option pour que XLSTAT adapte automatiquement le minimum et le maximum de l'axe des abscisses et de l'axe des ordonnées, une fois les transformations effectuées, de telle sorte que le graphique soit orthonormé.

Fusion de graphiques

Utilisez cet outil pour fusionner plusieurs graphiques en un seul.

Dans cette section :

Boîte de dialogue

Important : avant de lancer cet outil, vous devez sélectionner au moins deux graphiques du même type (par exemple, deux graphiques nuages de points).

OK : cliquez sur ce bouton pour fusionner les graphiques.

Annuler : cliquez sur ce bouton pour fermer la boîte de dialogue.

Aide : cliquez sur ce bouton pour afficher l'aide.

 : cliquez sur ce bouton pour rétablir les options par défaut.

 : cliquez sur ce bouton pour effacer les sélections de données.

Afficher le titre : activez cette option pour afficher un titre sur le graphique fusionné.

- **Titre du premier graphique** : activez cette option pour utiliser le titre du premier graphique.
- **Nouveau titre** : activez cette option pour entrer le titre du graphique fusionné.

Graphique orthonormé : activez cette option pour que XLSTAT vérifie après la fusion des graphiques que le graphique résultant est bien orthonormé.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Nouvelle feuille graphique : activez cette option pour afficher le graphique issu de la fusion des graphiques dans une nouvelle feuille graphique.

Afficher l'en-tête du rapport : désactivez cette option pour ne pas afficher l'en-tête du rapport précédant le graphique.

Analyse des données

Analyse factorielle

L'analyse factorielle (*factor analysis* en anglais), aussi appelée analyse factorielle des variables latentes, permet de mettre en évidence, lorsque cela est possible, l'existence de facteurs sous-jacents communs aux variables quantitatives mesurées pour un ensemble d'observations.

Dans cette section :

Description

La méthode de l'analyse factorielle date du début du 20^{ième} siècle (Spearman, 1904) et a connu de nombreux développements, plusieurs méthodes de calcul ayant été proposées. Si cette méthode a d'abord été utilisée par les psychométriciens, son champ d'application s'est peu à peu étendu à de nombreux autres domaines, par exemple en géologie, médecine, finance.

On distingue aujourd'hui deux grands types d'analyse factorielle :

- l'analyse factorielle exploratoire (en anglais, *exploratory factor analysis* ou EFA)
- l'analyse factorielle confirmatoire (en anglais, *confirmatory factor analysis* ou CFA)

L'EFA correspond à ce qui est décrit ci-dessous et à ce qui est utilisé par XLSTAT. Il s'agit d'une méthode qui permet de découvrir l'existence éventuelle de facteurs sous-jacents synthétisant l'information contenue dans un plus grand nombre de variables mesurées. La structure liant les facteurs aux variables est inconnue a priori et seul éventuellement le nombre de facteurs est supposé.

La CFA dans sa version traditionnelle s'appuie sur un modèle identique à celui de l'EFA, mais la structure liant les facteurs sous-jacents aux variables mesurées est supposée connue. Une version plus récente de la CFA est liée aux modèles d'équations structurelles.

Passer de p variables à k facteurs

L'exemple historique de Speaman, même s'il a depuis fait l'objet de nombreuses critiques et améliorations, permet de bien comprendre le principe et l'utilité de la méthode. En analysant les corrélations entre les notes obtenues par des enfants dans différentes matières, Spearman a voulu faire l'hypothèse que les notes dépendaient finalement d'un seul facteur, l'intelligence, avec une partie résiduelle due à un effet individuel, culturel ou autre.

Ainsi la note obtenue par l'individu (i) dans une matière (j) peut s'écrire $x(i,j) = \mu + b(j)F + e(i,j)$, avec μ la note moyenne de l'échantillon étudié, et où F est le niveau d'intelligence de l'individu (le facteur sous-jacent) et e(i,j) le résidu.

En généralisant cette écriture à p matières (les variables d'entrée) et à k facteurs sous-jacents, on obtient le modèle suivant :

$$x = \mu + \Lambda f + u \tag{1}$$

où x est un vecteur de dimension (p x 1), μ est le vecteur moyen, Λ est la matrice (p x k) des coordonnées factorielles (loadings en anglais) et f et u sont des vecteurs aléatoires de dimensions respectives (k x 1) et (p x 1), que l'on suppose indépendants. Les éléments de f sont appelés facteurs communs, et ceux de u facteurs spécifiques.

Si l'on s'impose que la norme de f vaut 1, alors la matrice de covariance des variables d'entrée sur la base de l'expression (1) s'écrit

$$\Sigma = \Lambda\Lambda' + \Psi \tag{2}$$

Ainsi, la variance de chacune des variables peut être divisée en deux parties : la communalité (car provenant des facteurs communs),

$$h_i^2 = \sum_{j=1}^{k} \lambda_{ij}^2 \tag{3}$$

et Ψ_{ii} la variance spécifique ou variance unique (car spécifique à la variable en question).

On peut montrer que la méthode qui permet de calculer la matrice Λ, enjeu essentiel de l'analyse factorielle, est indépendante de l'échelle. Il est donc équivalent de travailler à partir de la matrice de covariance ou de la matrice de corrélation.

L'enjeu de l'analyse factorielle est de permettre de trouver les matrices Λ et Ψ, de telle sorte que l'équation (2) soit au moins approximativement vérifiée.

Remarque : l'analyse factorielle est parfois rapprochée de l'Analyse en Composantes Principales (ACP), car l'ACP est un cas particulier de l'analyse factorielle (cas où k le nombre de facteurs vaut p le nombre de variables). Néanmoins ces deux méthodes ne sont en général pas utilisées dans le même contexte. En effet, l'ACP est avant tout utilisée pour réduire le nombre de dimensions tout en maximisant la variabilité conservée, pour obtenir des facteurs indépendants (non corrélés), ou pour visualiser les données dans un espace à 2 ou trois dimensions. L'analyse factorielle est quant à elle utilisée pour identifier une structure latente, et

pour éventuellement réduire par la suite le nombre de variables mesurées si elles sont redondantes vis-à- vis des facteurs latents.

Extraction des facteurs

Trois méthodes d'extraction des facteurs latents sont proposées par XLSTAT :

Composantes principales : cette méthode est aussi celle utilisée en Analyse en Composantes Principales (ACP). Elle n'est proposée ici que dans un but de comparaison entre les résultats des trois méthodes, sachant que les résultats proposés dans le module dédié à l'ACP sont plus complets.

Facteurs principaux : cette méthode est probablement la plus utilisée. C'est une méthode itérative qui permet de faire converger progressivement les communalités. Les calculs sont interrompus dès que le changement maximum des communalités est en dessous d'un seuil donné, ou lorsqu'un nombre maximal d'itérations est atteint. Les communalités initiales peuvent être calculées suivant différentes méthodes.

Maximum de vraisemblance : cette méthode a d'abord été proposée par Lawley (1940). La proposition de l'utilisation de l'algorithme de Newton- Raphson (méthode itérative) date de Jennrich (1969). Elle a ensuite été améliorée et généralisée par Jöreskog (1977). Cette méthode fait l'hypothèse que les variables d'entrée suivent une distribution normale. Les communalités initiales sont calculées suivant la méthode proposée par Jöreskog (1977). Dans le cadre de cette méthode, un test d'ajustement est calculé. La statistique utilisée pour le test suit une loi du χ^2 à $(p-k)^2/2-(p+k)/2$ degrés de liberté, où p est le nombre de variables et k le nombre de facteurs.

Nombre de facteurs

La détermination du nombre de facteurs à retenir est l'un des enjeux de l'analyse factorielle. La méthode « automatique » proposée par XLSTAT est uniquement basée sur la décomposition spectrale de la matrice de corrélation et sur la détection d'un seuil à partir duquel l'apport d'information (au sens de la variabilité) n'est pas significatif.

Si la méthode du maximum de vraisemblance propose un test d'ajustement pour aider à déterminer quel est le bon nombre de facteurs principaux, pour la méthode des facteurs principaux les méthodes sont plus empiriques.

La règle de Kaiser-Guttman propose de ne retenir que les facteurs pour lesquels les valeurs propres associées sont supérieurs strictes à 1 (les calculs doivent alors être effectués sur la matrice des corrélations). Le *scree test* (Cattell, 1966) est fondé sur la courbe décroissante des valeurs propres. Le nombre de facteurs à retenir correspond au premier point d'inflexion détecté sur la courbe. Des méthodes de validation croisée ont aussi été proposées dans ce but.

Cas problématiques (*Heywood cases*)

Les communalités sont par définition des carrés de corrélations. Elles doivent donc être comprise entre 0 et 1. Néanmoins, il se peut que les algorithmes itératifs (méthode des facteurs

principaux ou du maximum de vraisemblance) engendrent des solutions pour lesquelles les communalités sont égales à 1 (*Heywood cases*) ou supérieures à 1 (*ultra Heywood cases*). Les raisons de telles anomalies peuvent être multiples (trop de facteurs, pas assez de facteurs, …). Lorsque de tels cas sont rencontrés XLSTAT fixe les communalités à 1 et adapte en conséquence les éléments de **L**.

Rotations

Une fois les résultats obtenus, il est possible de les transformer afin de les rendre plus facilement interprétables, par exemple en essayant de faire en sorte que les coordonnées des variables sur les facteurs soient ou élevées (en valeur absolue), ou proches de zéro. On distingue deux grandes familles de rotations :

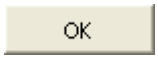
les rotations orthogonales peuvent être utilisées lorsque les facteurs ne sont pas corrélés (d'où orthogonales). Les méthodes proposées par XLSTAT sont Varimax, Quartimax, Equamax, Parsimax, Orthomax. La rotation Varimax est la plus utilisée. Elle permet de faire en sorte que pour chaque facteur, il y ait peu de coordonnées factorielles (loadings) élevées, et beaucoup de faibles. L'interprétation est ainsi facilitée puisqu'en principe les variables initiales seront surtout associées à l'un des facteurs.

les transformations obliques peuvent être utilisées lorsque les facteurs sont corrélés (d'où obliques). Les méthodes proposées par XLSTAT sont Quartimin et Oblimin.

La méthode Promax, également proposée par XLSTAT, est une procédure mixte puisqu'elle consiste d'abord en une rotation Varimax, puis en une rotation oblique telle que les coordonnées factorielles (*loadings*) élevées et faibles soient les mêmes, mais avec les valeurs faibles encore plus faibles.

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les

variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.

Onglet **Général**:

Le champ principal de saisie des données vous permet de sélectionner alternativement trois types de tableaux :

Tableau observations/variables / Matrice de corrélation / Matrice de covariance : choisissez l'option qui correspond au format de vos données, puis sélectionnez les données. Dans le cas de l'option **Tableau observations/variables** sélectionnez un tableau comprenant N observations décrites par P variables quantitatives. Dans le cas d'une **matrice de corrélation ou de covariance** sélectionnez une matrice carrée. Si des en- têtes de colonnes ont été sélectionnés, veuillez vérifier que l'option « Libellés des variables » est activée. Dans le cas d'une matrice de corrélation ou de covariance, si les libellés des colonnes sont sélectionnés, ceux des lignes doivent l'être aussi.

Corrélation : choisissez le type de matrice qui doit être utilisé par l'analyse factorielle. Le cas Pearson (n) se distingue du cas Pearson (n-1) par la façon dont sont normalisées les variables. Cela n'a d'influence que sur les coordonnées des observations.

Méthode d'extraction : choisissez la méthode d'extraction des facteurs à utiliser. Les trois méthodes possibles sont (voir la section description pour plus de détails) :

- Composantes principales
- Facteurs principaux
- Maximum de vraisemblance

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des variables : activez cette option si la première ligne des données sélectionnées (Tableau observations/variables, libellés des observations, poids) contient un libellé. Dans le cas où la sélection est une matrice de corrélation ou de covariance, si cette option est activée, la première colonne doit aussi comprendre le libellé des variables.

Libellés des observations : activez cette option si vous voulez utiliser des libellés d'observations pour l'affichage des résultats. Si l'option « Libellés des variables » est activée, la première cellule de la sélection doit comprendre un en-tête. Si vous n'activez pas cette option, des libellés seront automatiquement créés (Obs1, Obs2, …).

Poids : activez cette option si vous voulez pondérer les observations. Si vous n'activez pas cette option, les poids seront tous considérés comme valant 1. Les poids doivent être impérativement supérieurs ou égaux à 0. Si un en- tête de colonne a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

Onglet **Options**:

Nombre de facteurs :

- **Automatique** : activez cette option pour que XLSTAT détermine automatiquement le nombre de facteurs.
- **Défini par l'utilisateur** : activez cette option pour indiquer à XLSTAT quel est le nombre de facteurs à considérer pour les calculs.

Communalités initiales : choisissez la méthode de calcul des communalités initiales (cette option n'est visible que dans le cas de la méthode des facteurs principaux) :

- **Carrés des corrélations multiples** : les communalités initiales sont basées sur le niveau de dépendance d'une variable vis-à-vis des autres variables.
- **Aléatoires** : les communalités initiales sont tirées dans l'intervalle]0 ; 1[.
- **1** : les communalités initiales sont fixées à 1.
- **Maximum** : les communalités initiales sont fixées à la valeur maximum des carrés des corrélations multiples.

Conditions d'arrêt :

- **Itérations** : entrez le nombre maximal d'itérations pour l'algorithme. Les calculs sont interrompus dès que le nombre maximal d'itérations est dépassé. Valeur par défaut : 50.
- **Convergence** : entrez la valeur seuil d'évolution maximale des communalités d'une itération à l'autre, qui une fois atteinte permet de considérer que l'algorithme a convergé. Valeur par défaut : 0,0001.

Rotation : activez cette option si vous voulez appliquer une rotation à la matrice des coordonnées factorielles.

- **Nombre de facteurs** : entrez le nombre de facteurs auxquels la rotation doit être appliquée.
- **Méthode** : choisissez la méthode de rotation à utiliser. Pour certaines méthode la valeur d'un paramètre doit être entrée (Gamma pour Orthomax, Tau pour Oblimin, et la puissance pour Promax).

- **Normalisation de Kaiser** : activez cette option pour appliquer la normalisation de Kaiser pendant le calcul des rotations.

Onglet **Données manquantes** :

Ne pas accepter les valeurs manquantes : activez cette option pour que XLSTAT empêche la poursuite des calculs si des valeurs manquantes sont détectées.

Supprimer les observations : activez cette option pour supprimer les observations comportant des données manquantes.

Suppression par paire : activez cette option pour supprimer les observations comportant des données manquantes uniquement lorsque les variables impliquées dans les calculs comportent des données manquantes. Par exemple lors du calcul d'une corrélation entre deux variables, une observation ne sera ignorée que si la donnée correspondant à l'une des deux variables est manquante.

Estimer les données manquantes : activez cette option pour estimer les données manquantes avant le début des calculs.

- **Moyenne ou mode** : activez cette option pour estimer les données manquantes en utilisant la moyenne (variables quantitatives) ou le mode (variables qualitatives) pour les variables correspondantes.
- **Plus proche voisin** : activez cette option pour estimer les données manquantes d'une observation en recherchant le plus proche voisin de l'observation.

Onglet **Sorties**:

Statistiques descriptives : activez cette option pour afficher les statistiques descriptives pour les variables sélectionnées.

Corrélations : activez cette option pour afficher la matrice de corrélation ou de covariance en fonction du type d'options choisi dans l'onglet « Général ». Si l'option « **Tester la significativité** » est activée, les corrélations significatives au seuil de signification sont affichées en gras.

Kaiser-Meyer-Olkin : activez cette option pour calculer la statistique de la précision d'échantillonnage (*Measure of Sampling Adequacy* en anglais) de Kaiser-Meyer-Olkin.

Alpha de Cronbach : activez cette option pour calculer et afficher l'alpha de Cronbach.

Valeurs propres : activez cette option pour afficher le tableau et le graphique (scree plot) des valeurs propres.

Coordonnées factorielles : activez cette option pour afficher les coordonnées factorielles (coordonnées des variables dans l'espace des facteurs).

Corrélations Variables/ Facteurs : activez cette option pour afficher les corrélations entre les facteurs et les variables.

Coefficients du modèle factoriel : activez cette option pour afficher les coefficients du modèle factoriel. La multiplication des coordonnées (centrées et réduites) des observations dans l'espace d'origine par ces coefficients permet d'obtenir les coordonnées des observations dans l'espace des facteurs.

Structure factorielle : activez cette option pour afficher les corrélations entre les variables et les facteurs après rotation.

Onglet **Graphiques** :

Graphiques des variables : activez cette option pour afficher les graphiques de représentation des variables dans le nouvel espace.

- **Vecteurs** : activez cette option pour afficher les variables d'origine sous forme de vecteurs.

Graphiques de corrélations : activez cette option pour afficher les graphiques mettant en jeu des corrélations entre des composantes et des variables initiales.

- **Vecteurs** : activez cette option pour afficher les variables d'origine sous forme de vecteurs.

Graphiques des observations : activez cette option pour afficher les graphiques de représentation des observations dans le nouvel espace.

- **Etiquettes** : activez cette option pour afficher les étiquettes des observations sur les graphiques. Le nombre d'étiquettes affichées peut être modulé à l'aide de l'option de filtrage.

Etiquettes colorées : activez cette option pour que les étiquettes soient de la même couleur que les points correspondants.

Filtrer : activez cette option pour fixer le nombre d'observations affichées :

- **Aléatoire** : les observations à afficher sont sélectionnées de manière aléatoire. Le « Nombre d'observations » doit alors être saisi.
- **N premières lignes** : les N premières observations sont affichées. Le « Nombre d'observations » N doit alors être saisi.
- **N dernières lignes** : les N dernières observations sont affichées. Le « Nombre d'observations » N doit alors être saisi.
- **Variable de groupe** : si vous choisissez cette option, vous devez ensuite sélectionner une variable indicatrice composée de 1 pour les observations à afficher, et de 0 pour les observations à ne pas afficher.

Résultats

Statistiques descriptives : le tableau de statistiques descriptives présente pour toutes les variables sélectionnées des statistiques simples. Sont affichés le nombre d'observations, le nombre de données manquantes, le nombre de données non manquantes, la moyenne, et l'écart-type (non biaisé).

Matrice de corrélation /de covariance : ce tableau correspond aux données qui sont ensuite utilisées pour les calculs. Le type de corrélation dépend de l'option qui a été choisie dans l'onglet « Général » de la boîte de dialogue. Dans le cas de corrélations, les corrélations significatives sont affichées en gras.

Mesure de précision de l'échantillonnage de Kaiser-Meyer-Olkin : ce tableau donne pour chaque variable la valeur de la mesure KMO ainsi que le KMO global. L'indice KMO varie entre 0 et 1. Une valeur faible correspond au cas où il n'est pas possible d'extraire de facteurs synthétiques (ou variables latentes). Autrement dit, les individus ne permettent pas de faire ressortir le modèle que l'on pouvait imaginer préalablement (l'échantillon est « inadéquat »). Kaiser (1974) recommande de ne pas accepter une décomposition si le KMO est inférieur à 0.5. Si le KMO est entre 0.5 et 0.7 alors la qualité de l'échantillon est moyenne, elle est bonne pour un KMO entre 0.7 et 0.8, très bonne entre 0.8 et 0.9 et excellente au-delà.

Alpha de Cronbach : si l'option correspondante a été activée, la valeur du alpha de Cronbach est affichée.

Changement maximum de communalité à chaque itération : ce tableau permet d'observer l'évolution du changement maximum de communalité pour les 10 dernières itérations. Dans le cas de la méthode du maximum de vraisemblance, l'évolution d'un critère proportionnel à l'opposé du maximum de vraisemblance est aussi affichée.

Test d'ajustement : le test d'ajustement n'est affiché que dans le cas où la méthode du maximum de vraisemblance a été choisie.

Matrice des corrélations reproduites : cette matrice est le produit de la matrice des coordonnées factorielles par sa transposée.

Matrice de corrélation résiduelle : cette matrice est calculée comme la différence entre la matrice de corrélation des variables, et la matrice des corrélations reproduites.

Valeurs propres : dans ce tableau sont affichées les valeurs propres associées aux différents facteurs, ainsi que les pourcentages et pourcentages cumulés correspondants.

Vecteurs propres : dans ce tableau sont affichées les vecteurs propres.

Coordonnées factorielles : dans ce tableau sont affichées les coordonnées factorielles (coordonnées des variables dans l'espace des facteurs, appelées *factor loadings* ou *factor pattern* en anglais). Le graphique correspondant est affiché.

Corrélations Variables/Facteurs : dans ce tableau sont affichées les corrélations entre les facteurs et les variables.

Coefficients du modèle factoriel : dans ce tableau sont affichés les coefficients du modèle factoriel. La multiplication des coordonnées (centrées et réduites) des observations dans l'espace d'origine par ces coefficients permet d'obtenir les coordonnées des observations dans l'espace des facteurs.

Dans le cas où une rotation a été demandée, les résultats de la rotation sont affichés, avec en premier la **matrice de rotation** appliquée aux coordonnées des variables. Suivent ensuite les pourcentages modifiés de variabilité associés à chacun des axes concernés par la rotation. Dans les tableaux suivants sont affichées les coordonnées des variables et des observations après rotation.

Structure factorielle : dans ce tableau sont affichées les corrélations entre les variables et les facteurs après rotation.

Exemple

Un exemple d'utilisation de l'Analyse Factorielle est disponible sur le Centre d'aide XLSTAT à l'adresse

http://www.xlstat.com/demo-faf.htm

Bibliographie

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research*, **1**, 245-276.

Crawford C.B. and Ferguson G.A. (1970). A general rotation criterion and its use in orthogonal rotation. *Psychometrika*, **35(3)**, 321-332.

Cronbach L. J. (1951). Coefficient Alpha and the internal structure of test. *Psychometrika*, **16(3)**, 297-334.

Cureton E.E. and Mulaik S.A. (1975). The weighted Varimax rotation and the Promax rotation. *Psychometrika*, **40(2)**, 183-195.

Jennrich R.I. and Robinson S.M. (1969). A Newton-Raphson algorithm for maximum likelihood factor analysis. *Psychometrika*, **34(1)**, 111-123.

Jöreskog K.G. (1967). Some contributions to maximum likelihood factor analysis. *Psychometrika*, **32(4)**, 443-481.

Jö reskog K.G. (1977). Factor Analysis by Least-Squares and Maximum Likelihood Methods, in Statistical Methods for Digital Computers, eds. K. Enslein, A. Ralston, and H.S. Wilf. John Wiley and Sons, New York.

Kaiser H. F. (1974). An index of factorial simplicity. *Psychometrika*, **39**, 31-36.

Lawley D.N. (1940). The estimation of factor loadings by the method of maximum likelihood. *Proceedings of the Royal Society of Edinburgh*. **60**, 64-82.

Loehlin J.C. (1998). Latent Variable Models: an introduction to factor, path, and structural analysis, LEA, Mahwah.

Mardia K.V., Kent J.T. and Bibby J.M. (1979). Multivariate Analysis. Academic Press, London.

Spearman C. (1904). General intelligence, objectively determined and measured. *American Journal of Psychology*, **15**, 201-293.

Analyse en Composantes Principales (ACP)

Utilisez l'Analyse en Composantes Principales pour analyser un tableau observations/variables quantitatives ou une matrice de corrélation ou de covariance. Cette méthode permet

- d'étudier et visualiser les corrélations entre les variables,
- d'obtenir des facteurs non corrélés qui sont des combinaisons linéaires des variables de départ,
- de visualiser les observations dans un espace à deux ou trois dimensions.

Dans cette section :

Description

L'Analyse en Composantes Principales (ACP) est l'une des méthodes d'analyse de données multivariées les plus utilisées. Dès lors que l'on dispose d'un tableau de données quantitatives (continues ou discrètes) dans lequel n observations (des individus, des produits, …) sont décrites par p variables (des descripteurs, attributs, mesures, …), si p est assez élevé, il est impossible d'appréhender la structure des données et la proximité entre les observations en se contentant d'analyser des statistiques descriptives univariées ou même une matrice de corrélation.

Utilisations de l'ACP

Il existe plusieurs applications pour l'ACP, parmi lesquelles :

- l'étude et la visualisation des corrélations entre les variables, afin d'éventuellement limiter le nombre de variables à mesurer par la suite ;
- l'obtention de facteurs non corrélés qui sont des combinaisons linéaires des variables de départ, afin d'utiliser ces facteurs dans des méthodes de modélisation telles que la régression linéaire, la régression logistique ou l'analyse discriminante ;
- la visualisation des observations dans un espace à deux ou trois dimensions, afin d'identifier des groupes homogènes d'observations, ou au contraire des observations atypiques.

Principe de l'ACP

L'ACP peut être considérée comme une méthode de projection qui permet de projeter les observations depuis l'espace à p dimensions des p variables vers un espace à k dimensions ($k < p$) tel qu'un maximum d'information soit conservée (l'information est ici mesurée au travers de la variance totale du nuage de points) sur les premières dimensions. Si l'information associée aux 2 ou 3 premiers axes représente un pourcentage suffisant de la variabilité totale du nuage de points, on pourra représenter les observations sur un graphique à 2 ou 3 dimensions, facilitant ainsi grandement l'interprétation.

Corrélations ou covariance

L'ACP utilise une matrice indiquant le degré de similarité entre les variables pour calculer des matrices permettant la projection des variables dans le nouvel espace. Il est commun d'utiliser comme indice de similarité le coefficient de corrélation de Pearson, ou la covariance. La corrélation de Pearson et la covariance présentent l'avantage de donner des matrices semi-définies positives dont les propriétés sont utilisées en ACP. Néanmoins on peut envisager d'utiliser d'autres indices. XLSTAT propose d'utiliser la corrélation de Spearman car la matrice de corrélation Spearman est également semi définie positive (propriété indispensable pour l'ACP). Une ACP effectuée sur une matrice de corrélation de Spearman est équivalente à une ACP classique effectuée sur la matrice des rangs des données initiales. Lorsque vous réalisez une ACP de Spearman, vous pouvez choisir d'afficher la matrice des rangs des données initiales dans la feuille de résultats.

Classiquement, on utilise un coefficient de corrélation et non la covariance car l'utilisation du coefficient de corrélation permet de supprimer les effets d'échelle : ainsi une variable variant entre 0 et 1 ne pèse pas plus dans la projection qu'une variable variant entre 0 et 1000. Toutefois, dans certains domaines, lorsque les variables sont supposées être sur des échelles identiques, ou lorsque l'on veut que la variance des variables influe sur la construction des facteurs, on utilise la covariance.

Dans le cas où ne serait disponible qu'une matrice de similarité, et non un tableau observations/variables, ou dans le cas où vous voudriez utiliser un autre indice de similarité, vous pouvez réaliser une ACP en partant de la matrice de similarité. Les résultats obtenus ne concernent alors que les variables, aucune information sur les observations n'étant disponible.

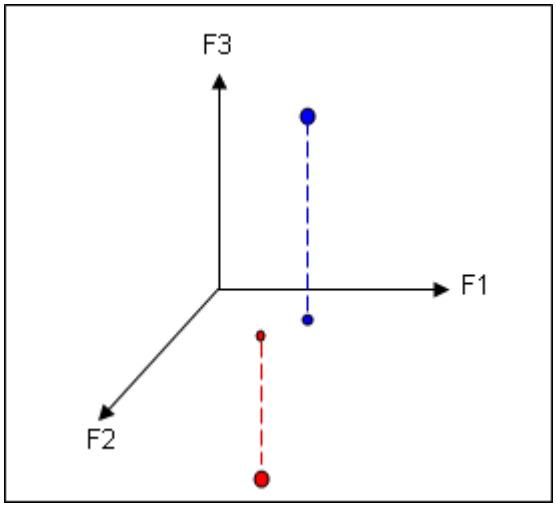
Remarque : dans le cas où l'ACP est réalisée sur une matrice de corrélation, on parle d'ACP normée.

Interprétation des résultats

La représentation des variables dans l'espace des k facteurs permet d'interpréter visuellement les corrélations entre les variables d'une part, et entre les variables et les facteurs d'autre part, moyennant certaines précautions.

En effet, qu'il s'agisse de la représentation des observations ou des variables dans l'espace des facteurs, deux points très éloignés dans un espace à k dimensions peuvent apparaître proches

dans un espace à 2 dimensions en fonction de la direction utilisée pour la projection (voir figure ci-dessous).



On peut considérer que la projection d'un point sur un axe, un plan ou un espace à 3 dimensions est fiable si la somme des cosinus carrés sur les axes de représentation n'est pas trop éloignée de 1. Les cosinus carrés sont affichés dans les résultats proposés par XLSTAT afin d'éviter toute mauvaise interprétation.

Si les facteurs doivent être utilisés par la suite avec d'autres méthodes, il est intéressant d'étudier la contribution relative (exprimée en % ou en proportion) des différentes variables à la construction de chacun des axes factoriels, afin de rendre les résultats obtenus ensuite facilement interprétables. Les contributions sont affichées dans les résultats proposés par XLSTAT.

Nombre de facteurs

Deux méthodes sont communément utilisées pour déterminer quel nombre de facteurs doit être retenu pour l'interprétation des résultats :

Le *scree test* (Cattell, 1966) est fondé sur la courbe décroissante des valeurs propres. Le nombre de facteurs à retenir correspond au premier point d'inflexion détecté sur la courbe.

On peut aussi se fonder sur le pourcentage cumulé de variabilité représenté par les axes factoriels et décider de se contenter d'un certain pourcentage.

Représentations graphiques

L'un des avantages de l'ACP est qu'elle fournit à la fois une visualisation optimale des variables et des observations, et des biplots mélangeant les deux (voir ci-dessous). Néanmoins, ces représentations ne sont fiables que si la somme des pourcentages de variabilité associés aux axes de l'espace de représentation, est suffisamment élevée. Si ce pourcentage est élevé (par exemple 80%), on peut considérer que la représentation est fiable. Si le pourcentage est faible, il est conseillé de faire des représentations sur plusieurs couples d'axes afin de valider l'interprétation faite sur les deux premiers axes factoriels.

Biplots

Suite à une ACP, il est possible de représenter simultanément dans l'espace des facteurs à la fois les observations et les variables. Les premiers travaux sur ce sujet datent de Gabriel (1971). Gower (1996) et Legendre (1998) ont synthétisé les travaux précédents et étendu cette technique de représentation graphique à d'autres méthodes. Le terme biplot est réservé aux représentations simultanées qui respectent le fait que la projection des observations sur les vecteurs variables doit être représentative des données d'entrée pour ces mêmes variables. Autrement dit, les points projetés sur le vecteur variable, doivent respecter l'ordre et les distances relatives des données de départ correspondant à la même variable.

La représentation simultanée des observations et des variables ne peut être faite directement en prenant les coordonnées des variables et des observations dans l'espace des facteurs. Une transformation est nécessaire afin de rendre l'interprétation exacte. Trois méthodes sont proposées en fonction du type d'interprétation que l'on souhaite pouvoir faire à partir de la représentation graphique :

- **biplot de corrélation** (*correlation biplot*) : ce type de biplot permet d'interpréter les angles entre les variables car ils sont directement liés aux corrélations entre les variables. La position de deux observations projetées sur un vecteur variable permet de conclure quant à leur niveau relatif sur cette même variable. La distance entre deux observations est une approximation de la distance de Mahalanobis dans l'espace des k facteurs. Enfin, la projection d'un vecteur variable dans l'espace de représentation est une approximation de l'écart-type de la variable (la longueur du vecteur dans l'espace des k facteurs est égale à l'écart-type de la variable).

- **biplot de distance** (*distance biplot*) : un biplot de distance permet d'interpréter les distances entre les observations car elles sont une approximation de leur distance euclidienne dans l'espace des p variables. La position de deux observations projetées sur un vecteur variable permet de conclure quant à leur niveau relatif sur cette même variable. Enfin, la longueur d'un vecteur variable dans l'espace de représentation est représentative du niveau de contribution de la variable à la construction de cet espace (la longueur du vecteur est la racine carrée de la somme des contributions).

- **biplot symétrique** (*symmetric biplot*) : ce biplot proposé par Jobson (1992) est intermédiaire entre les deux biplots précédents. Si ni les angles ni les distances ne peuvent être interprétés, on peut choisir cette représentation car elle est un compromis entre les deux.

XLSTAT vous donne la possibilité de jouer sur la longueur des vecteurs variables afin d'améliorer la lisibilité des graphiques. Néanmoins, si vous utilisez cette option dans le cas d'un biplot de corrélation, la projection d'un vecteur variable n'est plus une approximation de l'écart-type de la variable.

Graphiques bootstrap

Plusieurs techniques de validation des résultats existent en ACP. Un des objectifs de ces méthodes est d'évaluer la proximité entre les observations sur un plan factoriel et ainsi de savoir quelles observations sont significativement différentes les unes des autres. Pour cela,

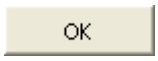
XLSTAT utilise la méthode du bootstrap partiel (Lebart, 2007). Le bootstrap partiel consiste à tirer m échantillons (avec remise) de même taille que la matrice de données utilisée pour l'ACP. Ensuite, chacun des échantillons est centré et potentiellement standardisé (dans le cas de l'ACP normée), puis les observations de chaque échantillon sont représentées sur les plans factoriels comme des observations supplémentaires. Ainsi, on dispose autour de chaque observation d'un nuage de points représentant les observations bootstrap. Afin d'alléger les graphiques, XLSTAT propose deux types de représentation :

- **Enveloppes convexes** : XLSTAT recherche les observations bootstrap les plus extrêmes et les relie entre elles afin de représenter l'enveloppe convexe du nuage de points.
- **Ellipses de confiance** : XLSTAT calcule et représente l'ellipse de confiance à 95% autour du nuage d'observations bootstrap lié à une observation donnée.

Ces deux représentations graphiques s'interprètent de la même manière. En effet, on pourra conclure que deux observations sont significativement différentes sur un plan factoriel donné si leurs enveloppes ou ellipses ne se chevauchent pas.

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.

Onglet **Général**:

Le champ principal de saisie des données vous permet de sélectionner alternativement trois types de tableaux :

Tableau observations/variables / Matrice de corrélation / Matrice de covariance : choisissez l'option qui correspond au format de vos données, puis sélectionnez les données.

Dans le cas de l'option **Tableau observations/variables** sélectionnez un tableau comprenant n observations décrites par p variables quantitatives. Dans le cas d'une **matrice de corrélation ou de covariance** sélectionnez une matrice carrée. Si des en- têtes de colonnes ont été sélectionnés, veuillez vérifier que l'option « Libellés des variables » est activée. Vous ne devez pas sélectionner les libellés des lignes.

Type d'ACP : Si le format observations/variables est sélectionné, vous avez le choix entre corrélation (ACP normée), covariance (ACP non normée) et Spearman pour effectuer l'ACP sur une matrice de corrélation de Spearman. Si le format des données est matrice de covariance, vous avez le choix entre corrélation (la matrice de covariance sélectionnée sera transformée en matrice de corrélation et on effectuera une ACP normée) ou covariance dans ce cas l'ACP sera réalisée sur la matrice de covariance sélectionnée et l'ACP ne sera pas normée.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des variables : activez cette option si la première ligne des données sélectionnées (Tableau observations/variables, libellés des observations, poids) contient un libellé.

Libellés des observations : activez cette option si vous voulez utiliser des libellés d'observations pour l'affichage des résultats. Si l'option « Libellés des variables » est activée, la première cellule de la sélection doit comprendre un en-tête. Si vous n'activez pas cette option, des libellés seront automatiquement créés (Obs1, Obs2, …).

Poids : activez cette option si vous voulez pondérer les observations. Si vous n'activez pas cette option, les poids seront tous considérés comme valant 1. Les poids doivent être impérativement supérieurs ou égaux à 0. Si un en- tête de colonne a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

Onglet **Options**:

Filtrer les facteurs : vous pouvez activer l'une des deux options suivantes afin de réduire le nombre de facteurs pour lesquels les résultats sont affichés :

- **% minimum** : activez cette option puis saisissez le pourcentage minimum de la variabilité totale que doivent représenter les facteurs retenus.
- **Nombre maximum** : activez cette option pour fixer le nombre maximum de facteurs à prendre en compte.

Normalisation : si le format des données est observations variables, vous pouvez choisir comment sont calculées les corrélations (ou covariance) : dénominateur (n) ou (n - 1).

Rotation : activez cette option si vous voulez appliquer une rotation à la matrice des coordonnées factorielles.

- **Nombre de facteurs** : entrez le nombre de facteurs pour lesquels la rotation sera appliquée.
- **Méthode** : choisissez la méthode de rotation à utiliser. Pour certaines méthode la valeur d'un paramètre doit être entrée (Kappa pour Orthomax, Tau pour Oblimin, et la puissance pour Promax).
- **Normalisation de Kaiser** : activez cette option pour appliquer la normalisation de Kaiser pendant le calcul des rotations.

Onglet **Données supplémentaires** :

Observations supplémentaires : activez cette option si vous voulez calculer les coordonnées et représenter des individus supplémentaires. Ces individus ne sont pas pris en compte pour le calcul des axes factoriels (observations passives, par opposition à observations actives). Si des libellés de variables sont présents pour les observations supplémentaires vous devez activer l'option « Libellés des variables pour les obs. ».

Variables supplémentaires : activez cette option si vous voulez calculer les coordonnées a posteriori pour des variables qui ne sont pas prises en compte pour le calcul des axes factoriels (variables passives, par opposition aux variables actives).

- **Quantitatives** : activez cette option si vous disposez de variables quantitatives supplémentaires. Si des en-têtes de colonnes ont été sélectionnés pour le tableau principal, veillez à ce qu'un libellé soit aussi présent pour les variables de cette sélection.
- **Qualitatives** : activez cette option si vous disposez de variables qualitatives supplémentaires. Si des en-têtes de colonnes ont été sélectionnés pour le tableau principal, veillez à ce qu'un libellé soit aussi présent pour les variables de cette sélection.
- **Afficher les barycentres** : activez cette option pour afficher les barycentres correspondant aux modalités des différentes variables qualitatives supplémentaires sélectionnées sur les graphiques des observations.

Onglet **Prétraitement** :

Données manquantes :

Ne pas accepter les valeurs manquantes : activez cette option pour que XLSTAT empêche la poursuite des calculs si des valeurs manquantes sont détectées.

Supprimer les observations : activez cette option pour supprimer les observations comportant des données manquantes.

Suppression par paires : activez cette option pour supprimer les observations comportant des données manquantes uniquement lorsque les variables impliquées dans les calculs comportent des données manquantes. Par exemple lors du calcul d'une corrélation entre deux variables, une observation ne sera ignorée que si la donnée correspondant à l'une des deux variables est manquante.

Estimer les données manquantes : activez cette option pour estimer les données manquantes avant le début des calculs.

- **Moyenne ou mode** : activez cette option pour estimer les données manquantes en utilisant la moyenne (variables quantitatives) ou le mode (variables qualitatives) pour les variables correspondantes.
- **Plus proche voisin** : activez cette option pour estimer les données manquantes d'une observation en recherchant le plus proche voisin de l'observation.

Remplacer les données manquantes par 0 : si le format des données est « matrice de corrélation » ou « matrice de covariance », vous pouvez activer cette option pour remplacer les données manquantes par 0.

Groupes :

Analyse par groupe : activez cette option puis sélectionnez ici les données d'appartenance à des groupes si vous souhaitez que les ACP soient effectués sur chaque groupe séparément. Vous pouvez choisir entre les options suivantes :

- **Une ACP par groupe** : Cette option permet de réaliser une ACP sur chacun des groupes.
- **Une ACP par groupe sélectionné** : Cette option vous permet de choisir à l'aide d'une boite de dialogue les groupes sur lesquels vous voulez réaliser les différentes ACP.
- **Une ACP sur un rassemblement de groupes** : Cette option vous permet de réaliser une ACP sur un rassemblement de groupes que vous sélectionnerez à l'aide d'une boite de dialogue.

Onglet **Sorties**:

Statistiques descriptives : activez cette option pour afficher les statistiques descriptives pour les variables sélectionnées.

Corrélations : activez cette option pour afficher la matrice de corrélation ou de covariance en fonction du type d'options choisi dans l'onglet « Général ».

- **Tester la significativité** : dans le cas où une corrélation a été choisie dans l'ongle « Général » de la boîte de dialogue, activez cette option pour tester la significativité des corrélations.
- **Test de sphéricité de Bartlett** : activez cette option pour effectuer le test de sphéricité de Bartlett.

- **Niveau de signification (%)** : entrez le niveau de signification pour les tests ci-dessus.
- **Kaiser-Meyer-Olkin** : activez cette option pour calculer la statistique de la précision d'échantillonnage (*Measure of Sampling Adequacy* en anglais) de Kaiser-Meyer-Olkin.

Matrice des rangs des données : Si vous avez choisis d'effectuer l'ACP sur la matrice de corrélation de Spearman, vous pouvez afficher la matrice des rangs des données brutes.

Valeurs propres : activez cette option pour afficher le tableau et le graphique (*scree plot*) des valeurs propres.

Coordonnées des variables : activez cette option pour afficher les coordonnées des variables dans l'espace des facteurs (*factor loadings* en anglais).

Corrélations Variables/Facteurs : activez cette option pour afficher les corrélations entre les facteurs et les variables.

Coordonnées des observations : activez cette option pour afficher les coordonnées des observations (*factor scores* en anglais) dans le nouvel espace créé par l'ACP.

Contributions : activez cette option pour afficher les tableaux des contributions pour les variables actives et les observations actives.

Cosinus carrés : activez cette option pour afficher les tableaux des cosinus carrés pour les variables et les observations.

Onglet **Graphiques** :

Sous-Onglet **Variables** :

Graphiques de corrélations : activez cette option pour afficher les graphiques mettant en jeu des corrélations entre des composantes et des variables initiales. Ce graphique est communément appelé cercle de corrélation.

- **Vecteurs** : activez cette option pour afficher les variables d'origine sous forme de vecteurs.
- **Orienter les libellés** : Cette option (disponible que sur les versions d'Excel supérieures à Excel 2010) permet d'orienter les libellés des variables de telle sorte qu'ils s'affichent dans la continuité du vecteur.
- **Etiquettes colorées** : activez cette option pour que les étiquettes soient de la même couleur que les points correspondants
- **Colorer par groupe** : activez cette option si vous souhaitez colorer les variables en fonction d'une variable de groupe, sélectionnez un vecteur colonne de taille égale au nombre de variables actives. Si des en-têtes de colonnes ont été sélectionnés pour le tableau principal, veillez à ce qu'un libellé soit aussi présent pour la variable de cette sélection.

- **Taille des points = f(Cos2)** : activez cette option si vous souhaitez que la taille des points correspondants aux variables soit proportionnelle à leur cosinus carré sur le plan sélectionné.

Filtrer : activez cette option pour filtrer les variables affichées :

- **Aléatoire** : les variables à afficher sont sélectionnées de manière aléatoire. Le « Nombre de variables » doit alors être saisi.
- **N premières variables** : les N premières variables sont affichées. Le « Nombre de variables » N doit alors être saisi.
- **N dernières variables** : les N dernières variables sont affichées. Le « Nombre de variables » N doit alors être saisi.
- **Variable de groupe** : si vous choisissez cette option, vous devez ensuite sélectionner une variable indicatrice composée de 1 pour les variables à afficher, et de 0 pour les variables à ne pas afficher.
- **Somme(Cos2)>** : choisissez cette option pour que, seules les variables ayant une somme des cosinus carrés supérieure à une valeur à saisir entre 0 et 1, soient affichées sur les graphiques de représentation des variables.

Sous-Onglet **Observations** :

Graphiques des observations : activez cette option pour afficher les graphiques de représentation des observations dans le nouvel espace.

- **Etiquettes** : activez cette option pour afficher les étiquettes des observations sur les graphiques. Le nombre d'étiquettes affichées peut être modulé à l'aide de l'option de filtrage.
- **Etiquettes colorées** : activez cette option pour que les étiquettes soient de la même couleur que les points correspondants
- **Colorer par groupe** : activez cette option si vous souhaitez colorer les observations en fonction d'une variable de groupe, sélectionnez un vecteur colonne de taille égale au nombre d'observations actives. Si des en-têtes de colonnes ont été sélectionnés pour le tableau principal, veillez à ce qu'un libellé soit aussi présent pour la variable de cette sélection.
- **Ellipses de confiance** : activez cette option si vous souhaitez afficher des ellipses de confiance autour des groupes d'observations correspondant aux différentes modalités de la variable de groupe sélectionnée pour colorer les observations. Vous devez également choisir un intervalle de confiance pour l'ellipse de confiance.
- **Taille des points = f(Cos2)** : activez cette option si vous souhaitez que la taille des points correspondants aux observations soit proportionnelle à leur cosinus carré sur le plan sélectionné.

Filtrer : activez cette option pour filtrer les observations affichées :

- **Aléatoire** : les variables à afficher sont sélectionnées de manière aléatoire. Le « Nombre de observations » doit alors être saisi.
- **N premières lignes** : les N premières variables sont affichées. Le « Nombre de observations » N doit alors être saisi.
- **N dernières lignes** : les N dernières variables sont affichées. Le « Nombre de observations » N doit alors être saisi.
- **Variable de groupe** : si vous choisissez cette option, vous devez ensuite sélectionner une variable indicatrice composée de 1 pour les observations à afficher, et de 0 pour les observations à ne pas afficher.

Somme(Cos2)> : choisissez cette option pour que, seules les observations ayant une somme des cosinus carrés supérieure à une valeur à saisir entre 0 et 1, soient affichées sur les graphiques de représentation des observations.

Sous-Onglet **Biplots** :

Biplots : activez cette option pour afficher les graphiques de représentation simultanée des observations et des variables d'origine dans le nouvel espace.

- **Options pour les variables** :
- **Vecteurs** : activez cette option pour afficher les variables d'origine sous forme de vecteurs.
- **Etiquettes** : activez cette option pour afficher les étiquettes des variables sur les biplots.
- **Options pour les observations** :
- **Etiquettes** : activez cette option pour afficher les étiquettes des observations sur les biplots.
- **Options communes** :
- **Obs/Var supp.** : si vous avez inclus des observations supplémentaires ou des variables supplémentaires dans l'analyse, activez cette option pour les afficher sur les biplots.
- **Filtrer Obs/Var** : si vous avez utilisé une variable de filtrage pour les observations ou pour les variables, cette même variable sera utilisée pour filtrer les observations et/ou les variables sur les biplots.
- **Colorer Obs/Var** : si vous avez utilisé une variable pour colorer les observations et/ou les variables, cette même variable de groupe sera utilisée pour colorer les observations et/ou les variables dans le biplot.

Type de biplots : choisissez le type de biplot que vous souhaitez afficher. Voir la section description pour plus de détails.

- **Biplot de corrélation** : activez cette option pour afficher des biplots de corrélation.
- **Biplot de distance** : activez cette option pour afficher des biplots de distance.
- **Biplot symétrique** : activez cette option pour afficher des biplots symétriques.
- **Coefficient** : choisissez le coefficient dont la racine carrée sera multipliée par les coordonnées des variables. Ce coefficient vous permettra d'ajuster la position des points variables dans le biplot afin de rendre ce dernier plus lisible. Si ce coefficient est différent de 1, la longueur des vecteurs variables n'est plus interprétable en termes d'écart-type (biplot de corrélation) ou de contribution (biplot de distance).

Sous-Onglet **Graphiques bootstrap**:

Graphique bootstrap des observations : activez cette option pour afficher les graphiques contenant les observations générées à l'aide de la méthode du bootstrap partiel. Voir la section description pour plus de détails.

- **Nombre d'échantillons bootstrap** : entrez ici le nombre d'échantillons bootsrap à générer.
- **Colorer les observations** : activez cette option pour que chaque observation soit colorée d'une couleur différente.
- **Filtrer les observations** : si vous avez utilisé une variable de filtrage pour les observations, cette même variable sera utilisée pour filtrer les observations sur le graphique bootstrap des observations.
- **Enveloppes convexes** : activez cette option pour que le nuage des observations bootstrap soit représenté au travers de son enveloppe convexe.
- **Ellipses de confiance** : activez cette option pour que le nuage des observations bootstrap soit représenté sous forme d'une ellipse de confiance.

Résultats

Statistiques descriptives : le tableau de statistiques descriptives présente pour toutes les variables sélectionnées des statistiques simples. Sont affichés le nombre d'observations, le nombre de données manquantes, le nombre de données non manquantes, la moyenne, et l'écart-type (non biaisé).

Matrice de corrélation/de covariance : ce tableau correspond aux données qui sont ensuite utilisées pour les calculs. Le type de corrélation dépend de l'option qui a été choisie dans l'onglet « Général » de la boîte de dialogue. Dans le cas de corrélations, les corrélations significatives sont affichées en gras.

Test de sphéricité de Bartlett : les résultats du test de sphéricité de Bartlett sont affichés. Ils permettent de valider ou d'infirmer l'hypothèse selon laquelle les variables ne sont pas significativement corrélées.

Mesure de précision de l'échantillonnage de Kaiser-Meyer-Olkin : ce tableau donne pour chaque variable la valeur de la mesure KMO ainsi que le KMO global. L'indice KMO varie entre 0 et 1. Une valeur faible correspond au cas où il n'est pas possible d'extraire de facteurs synthétiques (ou variables latentes). Autrement dit, les individus ne permettent pas de faire ressortir le modèle que l'on pouvait imaginer préalablement (l'échantillon est « inadéquat »). Kaiser (1974) recommande de ne pas accepter une décomposition si le KMO est inférieur à 0.5. Si le KMO est entre 0.5 et 0.7 alors la qualité de l'échantillon est moyenne, elle est bonne pour un KMO entre 0.7 et 0.8, très bonne entre 0.8 et 0.9 et excellente au-delà.

Valeurs propres : les valeurs propres et le graphique (*scree plot*) correspondant sont affichés. Le nombre de valeurs propres est égal au nombre de valeurs propres non nulles.

Si les options de sorties correspondantes ont été activées, XLSTAT affiche ensuite les **coordonnées des variables** dans le nouvel espace, puis les corrélations entre les variables d'origine et les composantes dans le nouvel espace. Les **corrélations** sont égales aux coordonnées des variables dans le cas d'une ACP normée (sur matrice de corrélation).

Si des variables supplémentaires ont été sélectionnées les coordonnées et les corrélations correspondantes sont affichées en fin de tableau.

Contributions : les contributions sont une aide à l'interprétation. Les variables ayant le plus influencé la construction des axes sont celles dont les contributions sont les plus élevées.

Indice d'homogénéité des axes: : Cet indice développé par nos équipes est très utiles pour déterminer si les contributions des observations sont homogènes pour les différents axes. Il est construit comme la proportion d'observations ayant une contribution absolue > 1/n. Un indice au dessus de 0.4 indique une très bonne homogénéité avec des observations bien représentées. En revanche, un indice inférieur à 0.1 doit être une alerte pour l'utilisateur qui devrait vérifier si il n'a pas de valeurs extrêmes sur les variables construisant l'axe qui fausseraient son interprétation (les valeurs extrêmes seraient alors les observations se démarquant des autres sur l'axe en question).

Cosinus carrés : comme pour les autres méthodes factorielles, l'analyse des cosinus carrés permet d'éviter des erreurs d'interprétation dues à des effets de projection. Si les cosinus carrés associés aux axes utilisés sur un graphique sont faibles, on évitera d'interpréter la position de l'observation ou de la variable en question.

Les **coordonnées des observations** dans le nouvel espace sont ensuite affichées. Si des données supplémentaires ont été sélectionnées, elles sont affichées en fin de tableau.

A la fin du tableau des coordonnées des observations, vous trouverez le bouton suivant :  .

Ce bouton vous permet d'ouvrir automatiquement la boite de dialogue pré-remplie de la CAH (Classification Ascendante Hiérarchique) afin d'effectuer une classification sur les coordonnées factorielles des observations.

Contributions : ce tableau fournit les contributions des observations à la construction des composantes principales.

Cosinus carrés : dans ce tableau sont affichés les cosinus carrés entre les vecteurs observations et les axes factoriels.

Dans le cas où une rotation a été demandée, les résultats de la rotation sont affichés, avec en premier la **matrice de rotation** appliquée aux coordonnées des variables. Suivent ensuite les pourcentages modifiés de variabilité associés à chacun des axes concernés par la rotation. Dans les tableaux suivants sont affichées les coordonnées, les contributions et les cosinus des variables et des observations après rotation.

Exemple

Un exemple d'utilisation de l'Analyse en Composantes Principales est disponible sur le Centre d'aide XLSTAT à l'adresse

http://www.xlstat.com/demo-pcaf.htm

Un exemple d'utilisation de l'Analyse en Composantes Principales avec filtrage sur les cosinus carrés est disponible sur le Centre d'aide XLSTAT à l'adresse

http://www.xlstat.com/demo-pcafilterf.htm

Bibliographie

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research*, **1**, 245-276.

Gabriel K.R. (1971). The biplot graphic display of matrices with application to principal component analysis. *Biometrika*, **58**, 453-467.

Gower J.C. and Hand D.J. (1996). Biplots. Chapman and Hall, London.

Jobson J.D. (1992). Applied multivariate data analysis. Volume II: Categorical and Multivariate Methods. Springer-Verlag, New York.

Jolliffe I.T. (2002). Principal Component Analysis, Second Edition. Springer, New York.

Kaiser H. F. (1974). An index of factorial simplicity. *Psychometrika*, **39**, 31-36.

Lebart L. (2007). Which bootstrap for principal axes methods? *Selected Contributions in Data Analysis and Classification.* P. Brito et al. Editors, Springer, 581-588.

Legendre P. and Legendre L. (1998). Numerical Ecology. Second English Edition. Elsevier, Amsterdam, 403-406.

Mauchly J. W. (1940). Significance test for sphericity of a normal n-variate distribution. The *Annals of Mathematical Statistics*. **11**, 204-209.

Morineau A. and Aluja-Banet T. (1998). Analyse en Composantes Principales. CISIA-CERESTA, Paris.

Rao C. R. (1964). The use and interpretation of principal components analysis in applied research. *Sankhya*, A **26**, 329-358.

Analyse factorielle de données mixtes (PCAmix)

Utilisez l'analyse factorielle de données mixtes (PCAmix) pour analyser un tableau de données où des observations sont décrites à la fois par des variables quantitatives et par des variables qualitatives. Cette méthode permet :

- d'étudier et visualiser les liens entre les variables,
- d'obtenir des facteurs non corrélés qui sont des combinaisons linéaires des variables de départ,
- de visualiser les observations dans un espace à deux ou trois dimensions.

Dans cette section :

Description

L'analyse factorielle de données mixtes est une méthode initiallement développées par Hill et Smith (1972). Différentes variantes de cette méthodes ont ensuite été développées (Escofier 1979, Pagès 2004). La méthode utilisée dans XLSTAT est la méthode appelée PCAmix développée par Chavent et al (2014). Cette méthode peut être vue comme un mélange de deux méthodes d'analyse factorielle bien connues : l'analyse en Composantes Principales (ACP) qui permet d'étudier un tableau observations/variables quantitatives et l'analyse des correspondances multiples (ACM) qui elle permet l'étude d'un tableau observations/variables qualitatives. La méthode PCAmix permet l'analyse d'un tableau de n observations décrites par p_1 variables quantitatives et par p_2 variables qualitatives ayant un total de m_Q modalités. On note $p = p_1 + p_2$. Comme les autres méthodes d'analyse factorielle, la méthode PCAmix permet de réduire la dimensionnalité des données et ainsi identifier les proximités entre les variables mais également les proximités entre les observations.

Résultats de PCAmix

La méthode PCAmix fournit les mêmes résultats classiques qu'une autre méthode d'analyse factorielle : les coordonnées factorielles, les contributions et les cosinus carrés. Ces résultats s'interprètent exactement de la même manière qu'en ACP ou en ACM. Les **contributions** sont une aide à l'interprétation. Les variables ayant le plus influencé la construction des axes sont celles dont les contributions sont les plus élevées. Les **cosinus carrés** permettent de mesurer la qualité de projection sur un axe factoriel et ainsi éviter des erreurs d'interprétation dues à des

effets de projection. Si les cosinus carrés associés aux axes utilisés sur un graphique sont faibles, on évitera d'interpréter la position de l'observation ou de la variable en question.

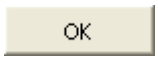
Une sortie spécifique à PCAmix est appelée les **"squared loadings"** des variables. Le "squared loading" entre une variable quantitative et un axe factoriel est égal à la corrélation au carré entre la variable et l'axe k. Le squared loading entre une variable qualitative y et un axe factoriel k est égal au rapport de corrélation η^2 entre la variable y et l'axe k. On a :

$$\eta^2(k|y) = \frac{\sum_{s=1}^{m} n_s(\bar{k_s} - \bar{k})^2}{\sum_{i=1}^{n}(k_i - \bar{k})^2}$$

où m est le nombre total de modalités de la variable y, n_s est le nombre d'observations possédant la modalité s, $\bar{k_s}$ est la moyenne de la variable k calculée sur les observations possédant la modalité s et $\bar{k}$ est la moyenne de la variable k calculée sur toutes les observations. Ce rapport de corrélation est égal à la somme des contributions à l'axe de chacune des s modalités de la variable qualitative y. Les squared loadings permettent de représenter sur un même graphique mixte les variables quantitatives et les variables qualitatives et ainsi voir leurs liens avec les axes factoriels.

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.

Onglet **Général**:

Observations/variables quantitatives : sélectionnez un tableau comprenant n observations décrites par p_1 variables quantitatives. Si des en-têtes de colonnes ont été sélectionnés,

veuillez vérifier que l'option "Libellés des variables" est activée. Vous ne devez pas sélectionner les libellés des lignes.

Observations/variables qualitatives : sélectionnez un tableau comprenant n observations décrites par p_2 variables qualitatives. Si des en-têtes de colonnes ont été sélectionnés, veuillez vérifier que l'option "Libellés des variables" est activée. Vous ne devez pas sélectionner les libellés des lignes.

Poids des observations : activez cette option si vous voulez pondérer les observations. Si vous n'activez pas cette option, les poids seront tous considérés comme valant 1. Les poids doivent être impérativement supérieurs ou égaux à 0. Si un en-tête de colonne a été sélectionné, veuillez vérifier que l'option "Libellés des variables" est activée.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des variables : activez cette option si la première ligne des données sélectionnées contient un libellé.

Libellés des observations : activez cette option si vous voulez utiliser des libellés d'observations pour l'affichage des résultats. Si l'option "Libellés des variables" est activée, la première cellule de la sélection doit comprendre un en-tête. Si vous n'activez pas cette option, des libellés seront automatiquement créés (Obs1, Obs2, …).

Onglet **Options**:

Tri alphabétique des modalités : activez cette option pour que dans les divers résultats, les modalités soient triées alphabétiquement pour chacune des variables qualitatives.

Libellés Variable-Modalité : activez cette option pour utiliser des libellés longs pour l'affichage des résultats. Les libellés Variable-Modalité sont composés du nom de la variable qualitative comme préfixe, et de la modalité comme suffixe.

Filtrer les facteurs : vous pouvez activer l'une des deux options suivantes afin de réduire le nombre de facteurs pour lesquels les résultats sont affichés :

- **% minimum** : activez cette option puis saisissez le pourcentage minimum de la variabilité totale que doivent représenter les facteurs retenus.
- **Nombre maximum** : activez cette option pour fixer le nombre maximum de facteurs à prendre en compte.

Onglet **Données supplémentaires** :

Observations supplémentaires : activez cette option si vous voulez calculer les coordonnées et représenter des individus supplémentaires. Ces individus ne sont pas pris en compte pour le calcul des axes factoriels (observations passives, par opposition à observations actives). Si des libellés de variables sont présents pour les observations supplémentaires vous devez activer l'option "Libellés des variables pour les obs. supp.". Vous pouvez également choisir des libellés des observations supplémentaires pour l'affichage des résultats.

- **Variables quantitatives** : sélectionnez un tableau contenant n' observations supplémentaires décrites par les p_1 variables quantitatives actives.
- **Variables qualitatives** : sélectionnez un tableau contenant n' observations supplémentaires décrites par les p_2 variables qualitatives actives.

Variables supplémentaires : activez cette option si vous voulez calculer les coordonnées a posteriori pour des variables qui ne sont pas prises en compte pour le calcul des axes factoriels (variables passives, par opposition aux variables actives).

- **Quantitatives** : activez cette option si vous disposez de variables quantitatives supplémentaires. Si des en-têtes de colonnes ont été sélectionnés pour les tableaux principaux, veillez à ce qu'un libellé soit aussi présent pour les variables de cette sélection.

- **Qualitatives** : activez cette option si vous disposez de variables qualitatives supplémentaires. Si des en-têtes de colonnes ont été sélectionnés pour les tableaux principaux, veillez à ce qu'un libellé soit aussi présent pour les variables de cette sélection.

Onglet **Données manquantes** :

Ne pas accepter les valeurs manquantes : activez cette option pour que XLSTAT empêche la poursuite des calculs si des valeurs manquantes sont détectées.

Supprimer les observations : activez cette option pour supprimer les observations comportant des données manquantes.

Remplacer les valeurs manquantes : activez cette option pour remplacer les valeurs manquantes. Pour les variables quantitatives, les données manquantes sont remplacées par la moyenne de la variable quantitative concernée, tandis que pour les variables qualitatives, une nouvelle catégorie « Manquant » est créée pour les variables qualitatives en question.

Onglet **Sorties**:

Statistiques descriptives : activez cette option pour afficher les statistiques descriptives pour les variables sélectionnées.

Valeurs propres : activez cette option pour afficher le tableau et le graphique (*scree plot*) des valeurs propres.

Affichez les résultats pour :

- **Observations et variables** : activez cette option pour afficher les résultats concernant les observations et les variables.
- **Observations** : activez cette option pour afficher uniquement les résultats concernant les observations.
- **Variables** : activez cette option pour afficher uniquement les résultats concernant les variables.

Coordonnées principales : activez cette option pour afficher les coordonnées principales.

Contributions : activez cette option pour afficher les contributions.

Cosinus carrés : activez cette option pour afficher les cosinus carrés.

Squared loadings : activez cette option pour afficher les squared loadings.

Onglet **Graphiques** :

Sous-Onglet **Quantitatives** :

Graphiques de corrélations : activez cette option pour afficher les graphiques mettant en jeu des corrélations entre des composantes et des variables initiales. Ce graphique est communément appelé cercle de corrélation.

- **Vecteurs** : activez cette option pour afficher les variables d'origine sous forme de vecteurs.
- **Orienter les libellés** : Cette option (disponible que sur les versions d'Excel superieures à Excel 2010) permet d'orienter les libellés des variables de telle sorte qu'ils s'affichent dans la continuité du vecteur.
- **Etiquettes colorées** : activez cette option pour que les étiquettes soient de la même couleur que les points correspondants.
- **Colorer par groupe** : activez cette option si vous souhaitez colorer les variables en fonction d'une variable de groupe, sélectionnez un vecteur colonne de taille égale au nombre de variables actives. Si des en-têtes de colonnes ont été sélectionnés pour le tableau principal, veillez à ce qu'un libellé soit aussi présent pour la variable de cette sélection.
- **Taille des points = f(Cos2)** : activez cette option si vous souhaitez que la taille des points correspondants aux variables soit proportionnelle à leur cosinus carré sur le plan sélectionné.

Filtrer : activez cette option pour filtrer les variables affichées :

- **Aléatoire** : les variables à afficher sont sélectionnées de manière aléatoire. Le « Nombre de variables » doit alors être saisi.
- **N premières variables** : les N premières variables sont affichées. Le « Nombre de variables » N doit alors être saisi.
- **N dernières variables** : les N dernières variables sont affichées. Le « Nombre de variables » N doit alors être saisi.
- **Variable de groupe** : si vous choisissez cette option, vous devez ensuite sélectionner une variable indicatrice composée de 1 pour les variables à afficher, et de 0 pour les variables à ne pas afficher.
- **Somme(Cos2)>** : choisissez cette option pour que, seules les variables ayant une somme des cosinus carrés supérieure à une valeur à saisir entre 0 et 1, soient affichées sur les graphiques de représentation des variables.

Sous-Onglet **Qualitatives**:

Carte factorielle des modalités : activez cette option pour afficher le graphique des coordonnées principales des modalités des variables qualitatives actives et supplémentaires.

- **Etiquettes** : activez cette option pour que les étiquettes des noms des modalités soient affichées sur le graphique.
- **Etiquettes colorées** : activez cette option pour que les étiquettes des noms des modalités soient de la même couleur que les points correspondants.
- **Colorer par groupe** : activez cette option si vous souhaitez colorer les variables en fonction d'une variable de groupe, sélectionnez un vecteur colonne de taille égale au nombre de variables actives. Si des en-têtes de colonnes ont été sélectionnés pour le tableau principal, veillez à ce qu'un libellé soit aussi présent pour la variable de cette sélection.
- **Taille des points = f(Cos2)** : activez cette option si vous souhaitez que la taille des points correspondants aux variables soit proportionnelle à leur cosinus carré sur le plan sélectionné.
- **Relier les modalités** : activez cette option pour que les modalités de chaque variable soient reliées entre elles. Cette option permet de repérer plus rapidement les modalités appartenant à une variable.

Filtrer : activez cette option pour filtrer les variables affichées :

- **Aléatoire** : les variables à afficher sont sélectionnées de manière aléatoire. Le « Nombre de variables » N doit alors être saisi.
- **N premières variables** : les N premières variables sont affichées. Le « Nombre de variables » N doit alors être saisi.
- **N dernières variables** : les N dernières variables sont affichées. Le « Nombre de variables » N doit alors être saisi.
- **Variable de groupe** : si vous choisissez cette option, vous devez ensuite sélectionner une variable indicatrice composée de 1 pour les variables à afficher, et de 0 pour les variables à ne pas afficher.

Sous-Onglet **Mixte** :

Graphique mixte : activez cette option pour afficher le graphique mixte représentant les squared loadings des variables quantitatives et qualitatives.

Options pour les variables quantitatives :

- **Etiquettes** : activez cette option pour afficher les étiquettes avec les noms des variables quantitatives.
- **Vecteurs** : activez cette option pour afficher des vecteurs reliant le centre du graphique et les points correspondants aux coordonnées des variables quantitatives.
- **Variables supplémentaires** : si vous avez inclus des variables supplémentaires quantitatives dans l'analyse, activez cette option pour les afficher sur le graphique mixte.
- **Filtrer les variables** : si vous avez utilisé une variable de filtrage pour les variables quantitatives, cette même variable sera utilisée pour filtrer les variables sur le graphique mixte.

Options pour les variables qualitatives :

- **Etiquettes** : activez cette option pour afficher les étiquettes avec les noms des variables qualitatives.
- **Vecteurs** : activez cette option pour afficher des vecteurs reliant le centre du graphique et les points correspondants aux coordonnées des variables qualitatives.
- **Variables supplémentaires** : si vous avez inclus des variables supplémentaires qualitatives dans l'analyse, activez cette option pour les afficher sur le graphique mixte.
- **Filtrer les variables** : si vous avez utilisé une variable de filtrage pour les variables qualitatives, cette même variable sera utilisée pour filtrer les variables sur le graphique mixte.

Sous-Onglet **Observations** :

Graphiques des observations : activez cette option pour afficher les graphiques de représentation des observations dans le nouvel espace.

- **Etiquettes** : activez cette option pour afficher les étiquettes des observations sur les graphiques. Le nombre d'étiquettes affichées peut être modulé à l'aide de l'option de filtrage.
- **Etiquettes colorées** : activez cette option pour que les étiquettes soient de la même couleur que les points correspondants.
- **Colorer par groupe** : activez cette option si vous souhaitez colorer les observations en fonction d'une variable de groupe, sélectionnez un vecteur colonne de taille égale au nombre d'observations actives. Si des en-têtes de colonnes ont été sélectionnés pour le tableau principal, veillez à ce qu'un libellé soit aussi présent pour la variable de cette sélection.
- **Taille des points = f(Cos2)** : activez cette option si vous souhaitez que la taille des points correspondants aux observations soit proportionnelle à leur cosinus carré sur le plan sélectionné.

Filtrer : activez cette option pour filtrer les observations affichées :

- **Aléatoire** : les variables à afficher sont sélectionnées de manière aléatoire. Le « Nombre d'observations » doit alors être saisi.
- **N premières lignes** : les N premières variables sont affichées. Le « Nombre d'observations » N doit alors être saisi.
- **N dernières lignes** : les N dernières variables sont affichées. Le « Nombre d'observations » N doit alors être saisi.
- **Variable de groupe** : si vous choisissez cette option, vous devez ensuite sélectionner une variable indicatrice composée de 1 pour les observations à afficher, et de 0 pour les observations à ne pas afficher.
- **Somme(Cos2)>** : choisissez cette option pour que, seules les observations ayant une somme des cosinus carrés supérieure à une valeur à saisir entre 0 et 1, soient affichées sur les graphiques de représentation des observations.

Onglet **Couleurs** :

Cet onglet vous permet de personnaliser les couleurs des différents résultats numériques et les couleurs des graphiques.

Résultats

Statistiques descriptives : le tableau de statistiques descriptives présente pour toutes les variables sélectionnées des statistiques simples.

Valeurs propres : les valeurs propres et le graphique (*scree plot*) correspondant sont affichés. Le nombre de valeurs propres est égal au nombre de valeurs propres non nulles.

Résultats pour les variables : Si les options de sorties correspondantes ont été activées, XLSTAT affiche ensuite les résultats pour les variables actives (coordonnées principales, cosinus carrés, contributions et squared loadings). La première partie de chaque tableau concerne les variables quantitatives, la seconde partie concerne les variables qualitatives. Si des variables supplémentaires ont été sélectionnées les résultats pour ces variables sont affichés ensuite. Le cercle des corrélations des variables quantitatives, la carte factorielle des modalités des variables qualitatives et le graphique mixte des squared loadings sont ensuite affichés.

Résultats pour les observations : Si les options de sorties correspondantes ont été activées, XLSTAT affiche ensuite les résultats pour les observations actives (coordonnées principales, cosinus carrés et contributions). Si des observations supplémentaires ont été sélectionnées les résultats pour ces observations sont affichés ensuite. La carte factorielle des observations est ensuite affichée.

*Remarque sur l'**indice d'homogénéité des axes*** : Cet indice développé par nos équipes est très utiles pour déterminer si les contributions des observations sont homogènes pour les différents axes. Il est construit comme la proportion d'observations ayant une contribution absolue > 1/n. Un indice au dessus de 0.4 indique une très bonne homogénéité avec des observations bien représentées. En revanche, un indice inférieur à 0.1 doit être une alerte pour l'utilisateur qui devrait vérifier si il n'a pas de valeurs extrêmes sur les variables construisant l'axe qui fausseraient son interprétation (les valeurs extrêmes seraient alors les observations se démarquant des autres sur l'axe en question).

Exemple

Un exemple d'utilisation de la méthode PCAmix est disponible sur le Centre d'aide XLSTAT à l'adresse

http://www.xlstat.com/demo-pcmf.htm

Bibliographie

Beh, E. J. and R. Lombardo (2012). A Genealogy of Correspondence Analysis. *Australian & New Zealand Journal of Statistics*. **54 (2)**, 137–168

Chavent, M., V. Kuentz, A. Labenne, B. Liquet, and J. Saracco (2014). PCAmixdata : Multivariate Analysis of Mixed Data. R package version 2.2.

Escofier, B. (1979). Traitement simultané de variables qualitatives et quantitatives en analyse factorielle. *Cahiers de l'analyse des données*. **4 (2)**, 137–146.

Hill, M. O. and A. J. E. Smith (1976, May). Principal Component Analysis of Taxonomic Data with Multi-State Discrete Characters. *Taxon*. **25 (2/3)**, 249–255.

Labenne, A. (2015). Méthodes de réduction de dimension pour la construction d'indicateurs de qualité de vie. Phd Thesis. Université de Bordeaux

Pagès, J. (2004). Analyse factorielle de données mixtes. *Revue de Statistique Appliquée*. **52(4)**, 93–111.

Analyse Factorielle Discriminante (AFD)

Utilisez l'analyse discriminante pour expliquer et prédire l'appartenance d'individus à plusieurs classes, sur la base de variables explicatives quantitatives ou qualitatives.

Dans cette section :

Description

L'Analyse Factorielle Discriminante (AFD) est une méthode ancienne (Fisher, 1936) qui dans sa version classique a peu évolué au cours des vingt dernières années. Cette méthode, à la fois explicative et prédictive, peut être utilisée pour :

- vérifier sur un graphique à deux ou trois dimensions si les groupes auxquels appartiennent les observations sont bien distincts,
- identifier quelles sont les caractéristiques des groupes sur la base de variables explicatives,
- prédire le groupe d'appartenance pour une nouvelle observation.

Les applications possibles de l'AFD sont très nombreuses, allant de l'écologie à la prévision de risque en finance (crédit scoring).

Le principe de l'AFD est de modéliser une variable dépendante qualitative Y à partir d'une nouvelle variable dite discriminante. Cette dernière est une combinaison linéaire des variables explicatives et est choisie de façon à ce qu'elle discrimine au mieux les classes définies par les modalités de la variable à expliquer.

Soient n individus décrits par p variables explicatives x_i = (x_i^1, ..., x_i^n), le but est de trouver une variable discriminante $s \in R^n$ combinaison linéraire des vecteurs x_1, ..., x_p telle que : $$ s = u_1 x_1 + . . . + u_p x_p , $$

où u = (u_1, . . . , u_p) $\in R^p$ est le vecteur des coefficients de cette combinaison linéaire, appelé le facteur discriminant. Pour cela il nous faut donc trouver d'une part le vecteur u et d'autre part une métrique pour justifier que s discrimine bien les classes.

Soit la matrice de variance-covariance totale T dont les éléments s'écrivent : $$ t ^ { j, l } = \frac{ 1 }{ n } \sum _ {i = 1}^{i = n} (x _ { i }^{ j } - \mu ^ { j })(x _ { i }^{ l } - \mu ^{ l }) , j, l = 1, ..., p,

\tag{1} $$

en supposant que tous les individus ont le même poids et où $$ \mu ^ {j} = \frac{ 1 }{ n } \sum_{i = 1}^{i = n} x_{ i }^{ j }, j= 1,..., p, $$

est la moyenne des observations sur la variable x_j. L'AFD repose sur l'analyse de la variance dont sa décomposition y est faite sur une partition des données de la façon suivante : $$ t ^ { j, l } = \frac{ 1 }{ n } \sum_{k = 1}^{k = g}\sum_{x_{ i } \in G_{ k }} (x _ { i }^{ j } - \mu_{ k } ^{ j } + \mu_{ k }^{ j } - \mu^{ j })(x _ { i }^{ l } - \mu_{ k }^{ l } + \mu_{ k }^{ l } - \mu ^{ l }) , j, l = 1, ..., p.\tag{2} $$

où μ_k est cette fois la moyenne de la classe G_k, pour k allant de 1 à g, le nombre de classes données par la variable dépendante. En développant le produit de l'équation (2) (voir Bardos (2001) pour plus de détails), on déduit que la matrice T se décompose en la somme de 2 matrices : B, la matrice de variance-covariance inter-classes des g centres de gravités μ_k, et W, la matrice de variance-covariance intra-classes résultant de la somme des matrices W_k obtenues pour chaque groupe G_k.

Une bonne discrimination des classes est possible si les centres de gravité projetés dans l'espace R^n sont bien éloignés et si les groupes projetés ne sont pas trop dispersés. Ceci est possible en maximisant le rapport entre la variance inter-groupe et la variance totale c'est-à-dire qu'on cherche, u tel que $$ \max_{u \in R^{ p }} \frac{u^{ ' }Bu}{u^{ ' }Tu}. $$

Par définition ce rapport est maximal si u est vecteur propre de $T^{-1}B$ associé à la plus grande valeur propre. Le nombre de valeurs propres non nulles est au plus égal à $(g-1)$ où g est le nombre de classes.

La fonction score permet par la suite de calculer pour chaque individu x son appartenance à une classe. Pour cela on se ramène au calcul des distances par rapport aux centres des groupes et à leur comparaison. La fonction s'écrit avec les notations ci-dessus : $$ f_ { i, k } (x) = \frac{1}{2} \delta_{i / k}(x) = (\mu _ { i } - \mu _ { k })^{ ' } M \left(x - \frac{\mu _ { i } + \mu _ { k }}{2}\right) , i, k = 1, ..., g\tag{3} $$

où M est une matrice symétrique définie positive et les $\left\{\left(\mu_i - \mu_k\right)^{'} M\right\}_{i,k}$ représentent les coefficients de la fonction discriminante canonique. On dispose alors de g distances, soit une par groupe. Pour un individu x l'ensemble des distances est calculé et comparé. L'individu x sera affecté au groupe qui donne la plus petite distance.

Matrices SSCP et matrices de distance

Dans le module AFD de XLSTAT on propose de calculer les matrices SSCP ou matrices des sommes de carrés et produits croisés. Elles se construisent comme les matrices de covariances et sont proportionnelles à celles-ci. Elles vérifient également la relation suivante : SSCP totale = SSCP inter + SSCP intra totale. En particulier la matrice SSCP intra est utilisée dans les calculs comme dans les tests statistiques ou le calcul des coefficients de la fonction discriminante canonique ($M = W^{-1}$).

Lorsque la matrice M de l'équation (3) est remplacée par l'inverse de la matrice des variances-covariances intra-classe W^{-1} alors on retombe sur la distance de Mahalanobis entre x et μ_i. Une description détaillée de ces matrices est faite dans l'aide ici. Dans le cas où l'on suppose

les matrices de variance-covariance intra-classes égales, la matrice des distances est calculée en utilisant la matrice de covariance intra-classes totale.

Dans le cas de l'hypothèse d'égalité des matrices de covariance, les distances de Fisher entre les classes sont calculées. Elles sont obtenues à partir de la distance de Mahalanobis et permettent un test de significativité. Dans le cas où l'on ne fait pas l'hypothèse d'égalité des matrices de covariance, les distances quadratiques généralisées entre les classes sont proposées dans les résultats. La distance généralisée est aussi calculée à partir des distances de Mahalanobis et tient compte des logarithmes des déterminants des matrices de covariance ainsi que des logarithmes des probabilités a priori.

Modèle linéaire ou quadratique

Deux modèles d'AFD sont possibles en fonction d'une hypothèse fondamentale : si l'on suppose que les matrices de covariance associées aux différentes classes de la variable dépendante sont identiques, on se trouve dans le cas de l'Analyse Factorielle Discriminante linéaire. Si l'on suppose au contraire que les matrices de covariance sont différentes pour au moins deux groupes, alors on se trouve dans le cadre d'un modèle quadratique. Cette hypothèse fondamentale est proposée dans les options du module AFD de XLSTAT. En cas d'incapacité de choisir la bonne option, le test de Box vous est proposé dans les options de sorties pour tester cette hypothèse. L'approximation de Bartlett qui s'appuie sur une loi du χ^2 permet également de réaliser ce test. Pour connaître la valeur de ce test, le mieux est de commencer par réaliser une analyse linéaire, puis, en fonction des résultats du test de Box, éventuellement faire une analyse quadratique.

Problèmes de multicolinéarité

Dans le cas du modèle linéaire et encore plus dans le cas du modèle quadratique on peut faire face à des problèmes de variables ayant une variance nulle ou de multicolinéarité entre variables. XLSTAT a été programmé de manière à éviter ces problèmes. Les variables responsables de tels problèmes sont automatiquement ignorées soit pour l'ensemble des calculs, soit, dans le cas du modèle quadratique, pour les groupes pour lesquels les problèmes se posent. Les statistiques de multicolinéarité sont optionnellement affichées afin de vous permettre d'identifier les variables sources de problèmes.

Méthodes pas à pas

Comme pour la régression linéaire et logistique, des méthodes pas à pas efficaces ont été proposées. Elles ne sont toutefois utilisables que lorsque seules des variables quantitatives sont sélectionnées car les tests d'entrée et sortie de variables s'appuient sur une hypothèse de normalité des variables. La méthode *stepwise* (pas à pas progressive) permet d'obtenir un modèle performant évitant les variables qui n'apportent que peu d'information au modèle. Ces méthodes sont proposées dans l'onglet option de l'application AFD de XLSTAT.

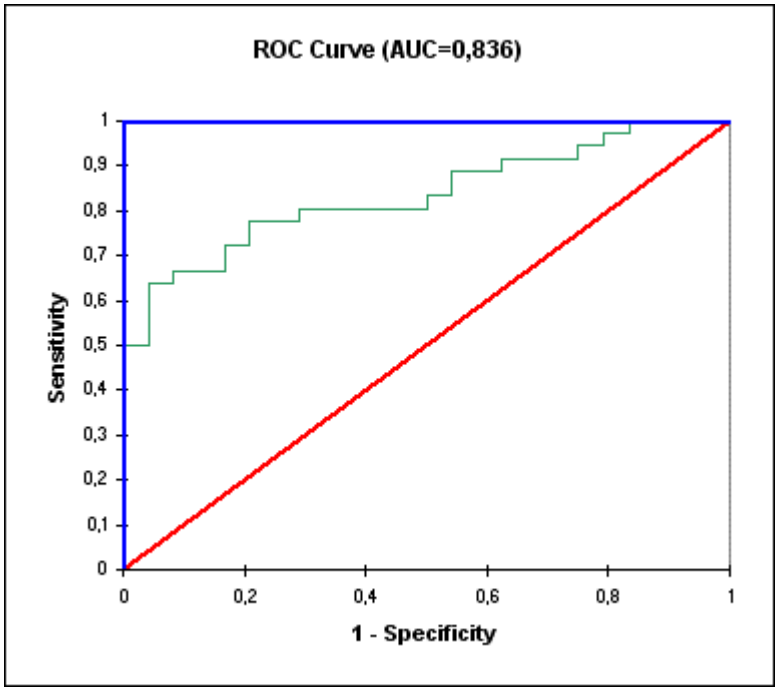
Tableau de classification, courbe ROC et validation croisée

Parmi les nombreux résultats proposés, XLSTAT donne la possibilité d'afficher le tableau de classification (aussi appelé matrice de confusion) qui permet de calculer un pourcentage

d'observations bien classées. Lorsque seules deux classes (ou catégories, ou modalités) sont présentes dans la variable dépendante, la courbe ROC peut aussi être affichée.

La courbe ROC permet de visualiser la performance d'un modèle, et de la comparer à celle d'autres modèles. La terminologie utilisée vient de la théorie de détection du signal. On désigne par sensibilité la proportion d'événements positifs bien classés. La spécificité correspond à la proportion d'événements négatifs bien classés. Si l'on fait varier la probabilité seuil à partir de laquelle on considère qu'un événement doit être considéré comme positif, la sensibilité et la spécificité varient. La courbe des points (1-spécificité, sensibilité) est la courbe ROC.

Considérons une variable dépendante binaire indiquant par exemple si un client a répondu favorablement à un mailing. Sur la figure ci-dessous, la courbe bleue correspond à un cas idéal où les $n\%$ de personnes ayant répondu favorablement correspondent aux $n\%$ de probabilités les plus élevées. La courbe verte correspond aux résultats d'un modèle bien discriminant. La courbe rouge (première bissectrice) correspond à ce que l'on obtiendrait avec un modèle aléatoire de Bernoulli avec une probabilité de réponse égale à celle observée sur l'échantillon étudié. Un modèle proche de la courbe rouge est donc inefficace puisqu'il n'est pas meilleur qu'un simple tirage au hasard. Un modèle en dessous de cette courbe serait catastrophique car il ferait moins bien que le hasard.



L'aire sous la courbe (ou AUC) est un indice synthétique calculé pour les courbes ROC. L'AUC correspond à la probabilité pour qu'un événement positif ait une probabilité donnée par le modèle plus élevée qu'un événement négatif. Pour un modèle idéal, on a $AUC = 1$, pour un modèle aléatoire, on a $AUC = 0.5$. On considère habituellement que le modèle est bon dès lors que la valeur de l'$AUC > 0.7$. Un modèle bien discriminant doit avoir une AUC entre 0.87 et 0.9. Un modèle ayant une $AUC > 0.9$ est excellent.

Les résultats du modèle en termes de prévision peuvent être trop optimistes : en effet, on cherche à vérifier si une observation est bien classée, alors qu'elle-même est prise en compte pour le calcul du modèle. Pour cette raison la validation croisée a été développée : pour

déterminer la probabilité d'appartenance d'une observation aux différentes classes, on la retire de l'échantillon d'apprentissage, puis on calcule le modèle et la prévision. Cette opération est répétée pour chacune des observations de l'échantillon d'apprentissage. Cette technique s'appelle le leave-one-out. Les résultats ainsi obtenus sont plus représentatifs de la qualité du modèle. XLSTAT propose de calculer les différentes statistiques associées à chacune des observations en mode validation croisée, ainsi que le tableau de classification et la courbe ROC s'il n'y a que deux classes.

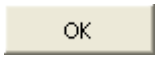
Enfin, il est conseillé de valider le modèle sur un échantillon de validation dans la mesure du possible. XLSTAT offre plusieurs possibilités pour automatiquement générer un échantillon de validation.

Analyse discriminante et régression logistique

Dans le cas où il n'y a que deux classes à prédire pour la variable dépendante, l'analyse discriminante est très proche de la régression logistique. L'analyse discriminante présente l'intérêt d'étudier dans le détail les structures de covariance, et d'aboutir à une représentation graphique. La régression logistique présente quant à elle l'avantage d'offrir plusieurs formes de modèles possibles, et de permettre l'utilisation des méthodes de sélection pas à pas y compris pour les variables explicatives qualitatives. L'utilisateur pourra comparer les performances des deux méthodes en s'appuyant sur les courbes ROC.

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.


: cliquez sur ce bouton pour changer la façon dont vous chargez les données dans XLSTAT. Si le bouton possède une icône de souris, XLSTAT vous permet de faire une sélection

à la souris des données. Si le bouton possède une icône de liste, XLSTAT vous permet de sélectionner les données à partir d'une liste. Si le bouton possède une icône orange, des boutons complémentaires avec un point d'interrogation  sont affichés pour vous permettre d'importer les données à partir de fichiers.

Onglet **Général** :

Y / Variables dépendantes :

Qualitatives : sélectionnez la ou les variables qualitatives que vous souhaitez modéliser. Si plusieurs variables sont sélectionnées, XLSTAT fera les calculs pour chacune des variables indépendamment. Si des en-têtes de colonnes ont été sélectionnés, veuillez vérifier que l'option « Libellés des variables » est activée.

X / Variables explicatives :

Quantitatives : activez cette option si vous voulez inclure une ou plusieurs variables explicatives quantitatives dans le modèle. Sélectionnez alors la ou les variables correspondantes sur la feuille Excel. Les données sélectionnées doivent être de type numérique. Si le libellé des variables a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

Qualitatives : activez cette option si vous voulez inclure une ou plusieurs variables explicatives qualitatives dans le modèle. Sélectionnez alors la ou les variables correspondantes sur la feuille Excel. Les données sélectionnées peuvent être de tout type, mais les données numériques sont automatiquement considérées comme nominales. Si le libellé des variables a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des variables : activez cette option si la première ligne des données sélectionnées (variables dépendantes et explicatives, libellés des observations, poids) contient un libellé.

Libellés des observations : activez cette option si vous voulez utiliser des libellés d'observations pour l'affichage des résultats. Si l'option « Libellés des variables » est activée, la première cellule de la sélection doit comprendre un en-tête. Si vous n'activez pas cette option, des libellés seront automatiquement créés (Obs1, Obs2, …).

Poids des observations : activez cette option si vous voulez pondérer les observations. Si vous n'activez pas cette option, les poids seront tous considérés comme valant 1. XLSTAT prend en compte ces poids pour les calculs des degrés de libertés. Les poids doivent être

impérativement supérieurs ou égaux à 0. Si un en-tête de colonne a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

Onglet **Options** :

Tolérance : entrez la valeur de la tolérance seuil en deçà de laquelle une variable est automatiquement ignorée.

Égalité des matrices de covariance : activez cette option si vous souhaitez faire l'hypothèse que les matrices de covariance associées aux différentes classes de la variable dépendante sont égales. Par conséquent, un modèle linéaire sera utilisé. Dans le cas contraire, ce sera le modèle quadratique qui sera choisi.

Probabilités a priori : activez cette option pour prendre en compte les probabilités a priori. Les probabilités associées à chacune des classes sont égales à la fréquence des classes. Remarque : cette option est sans effet si les probabilités a priori sont égales pour les différents groupes.

Filtrer les facteurs : vous pouvez activer l'une ou les deux options suivantes afin de réduire le nombre de facteurs pour lesquels les résultats sont affichés :

- **% minimum** : activez cette option puis saisissez le pourcentage minimum de la variabilité totale que doivent représenter les facteurs retenus.
- **Nombre maximum** : activez cette option pour fixer le nombre maximum de facteurs à prendre en compte.

Niveau de signification (%) : entrez le niveau de signification pour les différents tests calculés.

Sélection du modèle : activez cette option si vous souhaitez utiliser l'une des quatre méthodes de sélection proposées :

- **Stepwise (Ascendante)** : le processus de sélection commence par l'ajout de la variable ayant la plus forte contribution au modèle. Si une seconde variable est telle que sa probabilité d'entrée est supérieure à la **valeur seuil pour entrer**, alors elle est ajoutée au modèle. À partir de l'ajout de la troisième variable, après chaque ajout, on évalue pour toutes les variables présentes dans le modèle quel serait l'impact de son retrait. Si la probabilité de la statistique calculée est supérieure à la **valeur seuil pour retirer**, la variable est retirée du modèle.
- **Stepwise (Descendante)** : cette méthode est similaire à la précédente, mais part d'un modèle complet.
- **Ascendante** : la procédure est identique à celle de la sélection progressive, hormis le fait que les variables sont uniquement ajoutées et jamais retirées.
- **Descendante** : la procédure commence par l'ajout simultané de toutes les variables. Les variables sont ensuite retirées du modèle suivant la procédure utilisée pour la sélection progressive.

Correction du poids des classes : si les effectifs des différentes classes de la variable dépendante ne sont pas homogènes, on risque de pénaliser dans l'établissement du modèle les classes ayant un faible effectif. Pour palier ce problème, XLSTAT propose deux options :

- **Automatique** : le redressement est automatique. Des poids artificiels sont affectés aux observations dans le but d'obtenir des classes dont la somme des poids est identique.
- **Poids correctifs** : vous pouvez sélectionner les poids à affecter à chacune des observations.

Onglet **Validation** :

Validation : activez cette option si vous souhaitez utiliser une partie des données sélectionnées pour valider le modèle.

Échantillon de validation : choisissez l'une des options pour définir le mode de sélection des observations utilisées pour la validation :

- **Aléatoire** : les observations sont sélectionnées de manière aléatoire. Le « Nombre d'observations » doit alors être saisi.
- **N dernières lignes** : les N dernières observations sont sélectionnées pour la validation. Le « Nombre d'observations » N doit alors être saisi.
- **N premières lignes** : les N premières observations sont sélectionnées pour la validation. Le « Nombre d'observations » N doit alors être saisi.
- **Variable de groupe** : si vous choisissez cette option, vous devez ensuite sélectionner une variable indicatrice composée de 0 pour les observations à utiliser pour le calcul du modèle, et de 1 pour les observations à utiliser pour la validation du modèle.

Onglet **Prédiction** :

Prédiction : activez cette option si vous souhaitez sélectionner des données à utiliser en mode prédiction. Si vous activez cette option, vous devez veiller à ce que les données de prédiction soient organisées comme les données d'estimation : mêmes variables, même ordre dans les sélections. En revanche vous ne devez pas sélectionner de libellés de variables : la première ligne des sélections décrites ci-dessous doit être une ligne de données.

Quantitatives : activez cette option pour sélectionner la ou les variables quantitatives explicatives. La première ligne ne doit pas comprendre d'en-tête.

Qualitatives : activez cette option pour sélectionner la ou les variables qualitatives explicatives. La première ligne ne doit pas comprendre d'en-tête.

Libellés des observations : activez cette option si vous voulez utiliser des libellés d'observations disponibles sur une feuille Excel pour l'affichage des résultats. La première ligne

ne doit pas comprendre d'en-tête. Si vous n'activez pas cette option, des libellés seront automatiquement créés (PredObs1, PredObs2, …).

Onglet **Données manquantes** :

Supprimer les observations : activez cette option pour supprimer les observations comportant des données manquantes.

Estimer les données manquantes : activez cette option pour estimer les données manquantes avant le début des calculs.

- **Moyenne ou mode** : activez cette option pour estimer les données manquantes en utilisant la moyenne (variables quantitatives) ou le mode (variables qualitatives) pour les variables correspondantes.
- **Plus proche voisin** : activez cette option pour estimer les données manquantes d'une observation en recherchant le plus proche voisin de l'observation.

Onglet **Sorties** :

Sous-Onglet **Général** :

Statistiques descriptives : activez cette option pour afficher les statistiques descriptives pour les variables sélectionnées.

Corrélations : activez cette option pour afficher la matrice de corrélation.

Moyennes par classe : activez cette option pour afficher les moyennes de chaque classe pour chacune des variables explicatives.

Information sur chaque classe : activez cette option pour afficher des informations relatives à chaque classe à savoir la moyenne, la somme des poids, la probabilité d'affectation et le logarithme du déterminant calculé sur la matrice de covariance.

Statistiques de multicolinéarité : activez cette option pour afficher le tableau des statistiques de multicolinéarité.

Matrices SSCP : activez cette option pour afficher les matrices des sommes de carrés et produits croisés pour chaque variable explicative (intra-classe) et, éventuellement, chaque interaction (inter-classe) et totale.

Matrices de covariance : activez cette option pour afficher les matrices de covariance inter-classes, intra-classe, intra-classe totale, et totale.

Matrices de distance : activez cette option pour afficher les matrices des distances entre les groupes.

Valeurs propres : activez cette option pour afficher le tableau des valeurs propres.

Vecteurs propres : activez cette option pour afficher le tableau des vecteurs propres.

Corrélations Variables/Facteurs : activez cette option pour afficher les corrélations entre les facteurs et les variables.

Corrélations et fonctions canoniques : activez cette option pour les corrélations et les fonctions canoniques.

Fonctions de classement : activez cette option pour afficher les fonctions de classement.

Coordonnées des observations : activez cette option pour afficher les coordonnées des observations (*factor scores* en anglais) dans l'espace des facteurs. Dans ce tableau sont aussi affichées les classes a priori et a posteriori pour chaque observation, les probabilités d'affectation pour chaque classe, et les distances des observations à leur barycentre.

Matrice de confusion : activez cette option pour afficher le tableau permettant de visualiser les nombres d'observations bien et mal classées pour chacune des classes.

Validation croisée : activez cette option pour afficher les résultats concernant la validation croisée (probabilités pour les observations, et matrice de confusion).

Sous-Onglet **Test** :

Matrices de covariance intra-classe : activez cette option pour afficher les tests sur les matrices de covariance intra-classe.

- **Test de Box** : activez cette option pour que XLSTAT effectue le test de Box et les p-value découlant des deux approximations possibles.
- **Test de Kullback** : activez cette option pour que XLSTAT effectue le test de Kullback.

Moyennes par classe : activez cette option pour afficher les tests sur les moyennes des classes.

- **Test du lambda de Wilks (approximation de Rao)** : activez cette option pour que XLSTAT calcule la statistique Lambda et la p-value associée.
- **Trace de Pillai** : Activez cette option si vous voulez les résultats du test de la trace de Pillai.
- **Trace de Hotelling-Lawley** : activez cette option si vous voulez les résultats du test de la trace de Hotelling-Lawley.
- **Plus grande racine de Roy** : activez cette option si vous voulez les résultats du test de la plus grande racine de Roy.

Valeurs propres : activez cette option pour afficher le test sur les valeurs propres.

- **Test de Bartlett** : activez cette option si vous voulez les résultats du test de Bartlett. Ce test est possible que si le nombre de valeurs propres est strictement plus grand que 1.

Onglet **Graphiques** :

Graphiques de corrélations : activez cette option pour afficher le graphique mettant en jeu des corrélations entre les composantes et les variables initiales.

Valeurs propres : activez cette option pour afficher le graphique des valeurs propres (scree plot).

Graphiques des observations : activez cette option pour afficher les graphiques de représentation des observations dans l'espace des vecteurs propres.

- **Étiquettes** : activez cette option pour afficher les étiquettes des observations sur les graphiques. Le nombre d'étiquettes affichées peut être modulé à l'aide de l'option de filtrage.
- **Afficher les barycentres** : activez cette option pour afficher les barycentres correspondant aux modalités de la variable dépendante.
- **Ellipses de confiance** : activez cette option pour afficher des ellipses de confiance. Les ellipses de confiance correspondent à un intervalle de confiance à $x\%$ (x est déterminé à partir du niveau de signification spécifié dans l'onglet général) pour une loi normale bivariée de mêmes moyennes et de même matrice de covariance que les données factorielles correspondant aux différentes modalités de la variable dépendante.

Barycentres et cercles : activez cette option pour afficher le graphique des barycentres avec les cercles de confiance autour des moyennes.

Courbe ROC : activez cette option pour afficher le graphique de la courbe ROC, possible lorsque le nombre de groupes est exactement 2.

Étiquettes colorées : activez cette option pour que les étiquettes soient de la même couleur que les points correspondants.

Filtrer : activez cette option pour fixer le nombre d'observations affichées :

- **Aléatoire** : les observations à afficher sont sélectionnées de manière aléatoire. Le « Nombre d'observations » doit alors être saisi.
- **N premières lignes** : les N premières observations sont affichées. Le « Nombre d'observations » N doit alors être saisi.
- **N dernières lignes** : les N dernières observations sont affichées. Le « Nombre d'observations » N doit alors être saisi.
- **Variable de groupe** : si vous choisissez cette option, vous devez ensuite sélectionner une variable indicatrice composée de 1 pour les observations à afficher, et de 0 pour les observations à ne pas afficher.

Résultats

Statistiques descriptives : le tableau de statistiques descriptives présente pour toutes les variables sélectionnées des statistiques simples. Sont affichés le nombre d'observations, le

nombre de données manquantes, le nombre de données non manquantes, la moyenne, et l'écart-type (non biaisé).

Matrice de corrélation : ce résultat fournit tableau sont affichées les corrélations entre les variables explicatives.

Moyennes par classe : ce tableau fournit les moyennes des différentes variables explicatives pour les différentes classes de la variable dépendante.

Test unidimensionnel d'égalité des moyennes des classes : ce tableau fournit les résultats du test unidimensionnel d'égalité des moyennes des classes qui teste, variable par variable, l'hypothèse d'égalité des moyennes entre les classes. Le lambda de Wilks univarié est toujours compris entre 0 et 1. Une valeur de 1 correspond au cas où les moyennes des classes sont égales. Une valeur faible s'interprète comme de faibles variations intra-classe et donc de fortes variations inter-classes, d'où une différence significative des moyennes des classes.

Test du Lambda de Wilks (approximation de Rao) : ce tableau fournit les résultats du test du Lambda de Wilks qui teste l'hypothèse d'égalité des vecteurs moyens des différentes classes. Si lorsqu'il y a deux classes le test est équivalent au test de Fisher mentionné ci-après. Si le nombre de classes est inférieur ou égal à trois, le test est exact. L'approximation de Rao est nécessaire à partir de quatre classes pour obtenir une statistique approximativement distribuée suivant une loi de Fisher. Une description de ce test est faite dans l'aide ici.

Trace de Pillai : ce tableau fournit les résultats du test de Pillai qui teste l'hypothèse d'égalité des vecteurs moyens des différentes classes. Il est moins utilisé que le test du Lambda de Wilks et utilise aussi la loi de distribution de Fisher pour le calcul des p-values. Une description de ce test est faite dans l'aide ici.

Trace de Hotelling-Lawley : ce tableau fournit les résultats du test de la trace de Hotelling-Lawley qui teste l'hypothèse d'égalité des vecteurs moyens des différentes classes. Il est moins utilisé que le test du Lambda de Wilks et utilise aussi la loi de distribution de Fisher pour le calcul des p-values. Une description de ce test est faite dans l'aide ici.

Plus grande racine de Roy : ce tableau fournit les résultats du test de la plus grande racine de Roy qui teste l'hypothèse d'égalité des vecteurs moyens des différentes classes. Il est moins utilisé que le test du Lambda de Wilks et utilise aussi la loi de distribution de Fisher pour le calcul des p-values. Une description de ce test est faite dans l'aide ici.

Somme des poids, probabilités a priori et log des déterminants pour chaque classe : ce tableau fournit la somme des poids, la probabilité d'appartenance et le logarithme du déterminant pour chaque matrice de covariance. Ces statistiques sont utilisées entre autres dans les calculs des probabilités a posteriori pour les observations.

Statistiques de multicolinéarité : ce tableau permet d'identifier les variables responsables de multicolinéarités entre les variables. Dès qu'une variable est détectée comme étant responsable d'une multicolinéarité (sa tolérance est inférieure à la tolérance limite fixée dans l'onglet « options » de la boîte de dialogue), elle n'est pas prise en compte pour le calcul des statistiques de multicolinéarité des variables suivantes. Ainsi dans un cas extrême où deux variables seraient identiques, seule l'une des deux variables sera éliminée des calculs. Les statistiques affichées sont la tolérance (égale à $1 - R^2$), et son inverse, le VIF (Variance inflation factor).

Matrices SSCP : ce résultat fournit les matrices SSCP (*Sums of Squares and Cross Products*) inter, intra et totale l'une en dessous de l'autre.

Matrices de covariance : ce résultat fournit l'une en dessous de l'autre les matrices de variance-covariance inter-classes, intra-classe de chacun des groupes et totale.

Synthèse de la sélection des variables : ce tableau fournit une synthèse de la sélection des variables dans le cas où une méthode de sélection a été choisie dans les options. Dans le cas d'une sélection pas à pas (*stepwise*), Ascendante ou Descendante, les statistiques correspondant aux différentes étapes sont affichées.

Test de Box : ce tableau fournit les résultats du test de Box. Deux approximations ont été proposées, l'une basée sur la distribution du χ^2, l'autre sur la distribution de Fisher.

Test de Kullback : ce tableau fournit les résultats du test de Kullback. La statistique calculée est approximativement distribuée suivant une loi du χ^2.

Distances de Mahalanobis : ce résultat fournit les distances de Mahalanobis.

Distances de Fisher : ce résultat fournit les distances de Fisher. La matrice des p-values est également affichée afin de permettre de repérer quelles distances sont significatives.

Distances quadratiques généralisées : ce résultat fournit les distances quadratiques généralisées entre les groupes.

Valeurs propres : ce tableau fournit les valeurs propres associées aux différents facteurs, ainsi que les pourcentages et les pourcentages cumulés de discrimination correspondant. La somme des valeurs propres est égale à la trace de Hotelling. Le graphique *scree plot* est affiché en dessous si l'option a été choisie afin de visualiser comment le pouvoir discriminant est réparti entre les facteurs discriminants.

Vecteurs propres : ce tableau fournit les vecteurs propres qui interviennent ensuite dans le calcul des corrélations canoniques, des coefficients des fonctions canoniques et des coordonnées des observations.

Corrélations Variables/Facteurs : ce tableau fournit les résultats du calcul des corrélations entre les coordonnées des observations dans l'espace des variables initiales et dans l'espace des facteurs discriminants. Le graphique est affiché si l'option a été choisie pour permettre de visualiser sur un cercle des corrélations la relation entre les variables de départ et les facteurs. Le cercle des corrélations est une aide à l'interprétation de la représentation des observations dans l'espace des facteurs.

Corrélations canoniques : ce tableau fournit les corrélations canoniques associées à chaque facteur. Les corrélations canoniques sont aussi une mesure du pouvoir discriminant des facteurs. Leur somme est égale à la trace de Pillai.

Coefficients des fonctions discriminantes canoniques : ce tableau fournit les coefficients des fonctions discriminantes canoniques ils sont utilisés pour calculer les coordonnées d'une

observation dans l'espace des facteurs discriminants à partir de ses coordonnées dans l'espace des variables initiales.

Coefficients standardisés des fonctions discriminantes canoniques : ce tableau fournit les coefficients standardisés des coefficients donnés ci-dessus. Ainsi leur comparaison permet de mesurer la contribution relative des variables initiales à la discrimination pour un facteur donnée.

Fonctions aux barycentres : ce tableau fournit l'évaluation des fonctions discriminantes pour les points moyens pour chacune des classes.

Fonctions de classement : ce tableau fournit les fonctions de classement qui sont utilisées pour déterminer à quelle classe doit être affectée une observation sur la base des valeurs prises pour les différentes variables explicatives. Une observation est affectée à la classe pour laquelle la fonction de classement est la plus élevée.

Classification a priori, probabilités, coordonnées et carrés des distances : ce tableau fournit pour chaque observation, sa classe d'appartenance définie par la variable dépendante, la classe d'appartenance telle que déduite des probabilités d'appartenance, les probabilités d'appartenance à chacune des classes, les coordonnées dans l'espace des facteurs discriminants, et les carrés des distances des observations aux barycentres de chacune des classes.

N.B. : dans le cas où l'on ne fait pas l'hypothèse d'égalité des matrices de covariance, c'est la distance généralisée qui est affichée.

Matrice de confusion pour l'échantillon d'estimation : ce tableau fournit la matrice de confusion, ainsi que le pourcentage global d'observations biens classées. Dans le cas où la variable dépendante ne comprend que deux classes la courbe ROC est affichée (voir la section description pour plus détails).

Validation croisée : ce résultat fournit, dans le cas où l'option une validation croisée a été choisie, les informations pour les observations et la matrice de confusion (voir la section description pour plus détails).

Exemple

Un exemple d'utilisation de l'Analyse Factorielle Discriminante est disponible sur le Centre d'aide XLSTAT à l'adresse

http://www.xlstat.com/demo-daf.htm

Bibliographie

Bardos M. (2001). Analyse discriminante. Application au risque et scoring financier. Dunod, Paris.

Fisher R.A. (1936). The Use of Multiple Measurements in Taxonomic Problems. *Annals of Eugenics*, **7**, 179 -188.

Huberty C. J. (1994). Applied Discriminant Analysis. Wiley-Interscience, New York.

Jobson J.D. (1992). Applied multivariate data analysis. Volume II: Categorical and Multivariate Methods. Springer-Verlag, New York.

Lachenbruch P. A. (1975). Discriminant Analysis. Hafner, New York.

Tomassone R., Danzart M, Daudin J.J., Masson J.P. (1988). Discrimination et Classement. Masson, Paris.

Analyse Factorielle des Correspondances (AFC)

Utilisez ce module pour représenter graphiquement les proximités entre les modalités (aussi appelées catégories) de deux variables qualitatives. Les variables qualitatives peuvent être disponibles sous forme d'un tableau individus/variables, ou sous forme d'un tableau de contingence (tableau croisé).

Dans cette section :

Description

L'**Analyse Factorielle des Correspondances** (AFC) est une méthode qui permet d'étudier l'association entre deux variables qualitatives. Les travaux de J.-P. Benzécri commencés au début des années 60 ont permis l'émergence de la méthode. Ses disciples ont ensuite permis différentes évolutions. On citera notamment les contributions de M.J. Greenacre (1984) qui ont permis de généraliser l'approche et de la diffuser dans le monde anglo-saxon, et les travaux de C. Lauro qui a notamment mis au point une variante non symétrique de la méthode.

La mesure de l'association entre deux variables qualitatives est un sujet complexe qui nécessite une transformation préalable des données : en effet, il n'est pas possible de calculer un coefficient de corrélation en utilisant directement les données, comme on pourrait le faire avec deux variables quantitatives.

La première transformation consiste en un recodage des deux variables qualitatives V_1 et V_2 en deux tableaux disjonctifs Z_1 et Z_2. Pour chaque modalité de la variable V_j, une colonne est créée dans Z_j. A chaque fois qu'une modalité m de la variable V_j correspond à un individu i, on affecte 1 à $Z_1(i,m)$. Les autres valeurs de Z_1 et Z_2 sont nulles. La généralisation de cette idée à plus de deux variables correspond à l'Analyse des Correspondances Multiples. Lorsqu'il n'y a que deux variables, il est suffisant d'étudier le tableau de contingence des variables, qui n'est autre que le produit $Z_1^t Z_2$.

Un tableau de contingence a la structure suivante :

$V_1 \setminus V_2$	**Modalité 1**	…	**Modalité j**	…	**Modalité** m_2
Modalité 1	n_{11}		n_{1j}	…	n_{1m_2}
…	…	…	…	…	…
Modalité i	n_{i1}	…	n_{ij}	…	n_{im_2}
…	…	…	…	…	…
Modalité m_1	$n_{m_1 1}$	…	$n_{m_1 j}$	…	$n_{m_1 m_2}$

où n_{ij} est la fréquence des observations présentant à la fois la caractéristique i pour la variable V_1, et la caractéristique j pour la variable V_2.

La distance du Khi² a été proposée pour mesurer la distance entre les modalités. La somme de ces distances pour l'ensemble des cases du tableau donne la statistique du Khi² qui suit asymptotiquement une loi du Khi² à $(m_1 - 1)(m_2 - 1)$ degrés de liberté. Cette statistique permet de tester l'hypothèse d'indépendance entre les lignes et les colonnes du tableau de contingence.

La notion d'inertie inspirée de la physique est utilisée en Analyse Factorielle des Correspondances. L'inertie d'un nuage de points est la moyenne pondérée des carrés des distances au centre de gravité. Dans le cas de l'AFC, l'inertie totale du nuage des modalités est donnée par :

$$\phi^2 = \frac{\chi^2}{n}$$

$$\text{avec } \chi^2 = \sum_{i=1}^{m_1}\sum_{j=1}^{m_2} \frac{(n_{ij} - \frac{n_{i.}n_{.j}}{n})^2}{\frac{n_{i.}n_{.j}}{n}},\ n_{i.} = \sum_{j=1}^{m_2} n_{ij} \text{ et } n_{.j} = \sum_{i=1}^{m_1} n_{ij}$$

et où n est la somme des fréquences du tableau de contingence. On voit ici que l'inertie totale est proportionnelle à la statistique du Khi² de Pearson mesurée sur le tableau de contingence.

L'AFC consiste à représenter un maximum de l'inertie totale sur le premier axe factoriel, un maximum de l'inertie résiduelle sur le second axe, et ainsi de suite jusqu'à la dernière dimension. On montre que le nombre de dimensions de l'espace de représentation est inférieur ou égal à $min(m_1, m_2) - 1$.

L'**Analyse Non Symétrique des Correspondances** (ANSC) proposée par Lauro et D'Ambra (1984) permet d'étudier l'association entre les lignes et les colonnes d'un tableau de contingence tout en introduisant la notion de dépendance entre les lignes et les colonnes, d'où l'asymétrie. L'exemple historique présenté par Lauro et D'Ambra consiste en l'étude d'un tableau de contingence contenant les fréquences de prescription de 6 médicaments pour 7 maladies, et ce pour 69 patients. On voit bien ici qu'il y a une dépendance des médicaments vis-à-vis de la maladie. Afin de prendre en compte cette dépendance l'indice tau de Goodman et Kruskal (1954) a été retenu. L'indice correspondant au cas où les lignes dépendent des colonnes est donné par :

$$\tau_{b/RC} = \frac{\sum_{j=1}^{m_2} \frac{n_{.j}}{n} \sum_{i=1}^{m_1} (\frac{n_{ij}}{n_{.j}} - \frac{n_{i.}}{n})^2}{1 - \sum_{i=1}^{m_1} (\frac{n_{i.}}{n})^2}$$

Comme pour l'inertie totale, il est possible de calculer un espace de représentation des modalités, tel que la proportion du tau de Goodman et Kruskal conservée soit maximisée sur les premiers axes.

Greenacre (1984) a mis au point une approche calculatoire utilisant la décomposition en valeurs singulières qui permet de traiter dans un même cadre mathématique ces deux méthodes (AFC et ANSC).

Une alternative à l'Analyse Factorielle des Correspondances utilisant la **distance de Hellinger** a été proposée par Rao (1995). La distance de Hellinger dépend seulement des profils de la paire de modalités concernées et ne dépend pas de la taille de l'échantillon sur lequel les profils sont estimés. L'approche utilisant la distance de Hellinger peut donc être une bonne alternative à l'AFC classique lorsque le profil moyen colonne n'est pas pertinent (par exemple, lorsque les colonnes représentent des populations d'individus classés selon les catégories de lignes) ou si certaines catégories ont de faibles fréquences. Les calculs suivent l'approche unifiée décrite par Cuadras et Cuadras (2008). La formule généralisée de l'inertie est la suivante :

$$\phi^2(\alpha, \beta) = \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} [(\frac{\frac{n_{ij}}{n}}{\frac{n_{i.}}{n} \frac{n_{.j}}{n}})^\alpha - 1]^2 \frac{n_{i.}}{n} (\frac{n_{.j}}{n})^{2\beta}$$

Remarques :

- Dans le cas de l'Analyse Factorielle des Correspondances utilisant la distance de Hellinger, on a $\alpha = \frac{1}{2}$ et $\beta = \frac{1}{2}$.
- Dans le cas de l'Analyse Factorielle des Correspondances classique, on a $\alpha = 1$ and $\beta = \frac{1}{2}$

L'**Analyse d'un sous-ensemble** de modalités (ou catégories), est une méthode très récemment mise au point par Greenacre et Pardo (2006), qui permet de focaliser l'étude sur quelques catégories uniquement, tout en prenant en compte toutes les données du tableau de contingence grâce au maintien des sommes marginales du tableau. Sur des tableaux de taille importante cela permet de découper l'analyse en plusieurs sous-analyses.

L'**Analyse des Correspondances Détendancée** (ACD) est une méthode proposée par Hill et Gauch (1980), principalement utilisée sur des données écologiques. L'objectif de cette méthode est de corriger deux inconvénients rencontrés lors de l'utilisation de l'AFC classique :

- Le premier est l'« effet d'arc », aussi appelé l'« effet de fer à cheval ». Cet effet se traduit par la présence d'une courbe en forme d'arc dans les représentations graphiques des coordonnées, et est le résultat d'une relation quadratique entre les coordonnées sur les axes représentés. Cependant, pour que les axes soient interprétables séparemmment, il faut qu'ils soient indépendants et non corrélés. L'étape de détendance permet de corriger cet effet. Pour cela, lors du calcul des coordonnées, chaque axe est divisé en segments, puis les coordonnées pour l'axe calculé sont ajustées en soustrayant la moyenne mobile de chaque segment. Ainsi, lors du calcul des coordonnées du deuxième axe, le premier

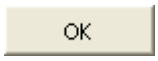
axe est découpé en segments et les coordonées calculées sont ajustées par rapport au premier axe. Pour les troisième et quatrième axes, ces étapes sont effectuées par rapport à chacun des axes précédemment obtenus.

- Le deuxième inconvénient souvent rencontré est la tendance que l'AFC a de comprimer les distances entre les points sur les extrémités des axes. L'Analyse des Correspondances Détendancée intégre alors une étape de remise à l'échelle non linéaire pour y remédier.

Remarque : L'ACD ne permet de calculer des coordonnées que sur un maximum de 4 dimensions.

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ces boutons pour changer la façon dont XLSTAT doit charger les données :

- cas où les données sont dans un tableau de contingence ou un tableau croisé : si la flèche est vers le bas, XLSTAT vous permet de sélectionner les données par colonnes ou par plage. Si la flèche est vers la droite, XLSTAT vous permet de sélectionner les données par lignes ou par plage ;

- cas où les données sont dans un tableau observations/variables : si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.


: cliquez sur ce bouton pour changer la façon dont vous chargez les données dans XLSTAT. Si le bouton possède une icône de souris, XLSTAT vous permet de faire une sélection à la souris des données. Si le bouton possède une icône de liste, XLSTAT vous permet de sélectionner les données à partir d'une liste. Si le bouton possède une icône orange, des

boutons complémentaires avec un point d'interrogation  sont affichés pour vous permettre d'importer les données à partir de fichiers.

Onglet **Général** :

Le champ principal de saisie des données vous permet de sélectionner alternativement deux types de tableaux :

Tableau croisé : choisissez cette option si vos données correspondent à un tableau croisé, avec dans chaque cellule les fréquences correspondant aux croisements des différentes catégories de deux variables qualitatives (dans ce cas on parle de tableau de contingence), ou des valeurs d'une autre nature.

Tableau observations/variables : choisissez cette option si vos données correspondent à un tableau comprenant N observations décrites par deux variables qualitatives. Ce type de tableau correspond typiquement à un questionnaire à deux questions. Ce tableau sera alors automatiquement transformé par XLSTAT en un tableau de contingence.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés inclus : cette option est visible si vous avez sélectionné un tableau de type tableau de contingence ou tableau de données. Activez cette option si vous avez inclus les libellés des lignes et des colonnes dans la sélection. Dans ce cas, la première colonne de la sélection contient les libellés des lignes et la première ligne contient les libellés des colonnes.

Libellés des variables : cette option est visible si vous avez sélectionné un tableau de type tableau observations/variables. Activez cette option si la première ligne de la sélection contient le libellé des variables.

Poids : cette option est visible si vous avez sélectionné un tableau de type tableau observations/variables. Activez cette option si vous voulez pondérer les observations. Si vous n'activez pas cette option, les poids seront tous considérés comme valant 1. Les poids doivent être impérativement supérieurs ou égaux à 0. Si un en-tête de colonne a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

Onglet **Options** :

Analyse approfondie : cette option permet de choisir le type d'analyse que vous souhaitez réaliser sur les données. L'analyse sur les données supplémentaires et l'analyse d'un sous-ensemble ne sont actives que dans le cas où les données sélectionnées correspondent à un tableau de contingence ou à un tableau croisé. Les options suivantes sont proposées :

- **Analyse détendancée** : si vous sélectionnez cette option vous pouvez ensuite entrer les paramètres utiles aux calculs, à savoir le nombre de segments pour découper les axes et le nombre de remises à l'échelle à effectuer. Par défaut, le nombre de segments est fixé à 26 et le nombre de remises à l'échelle est fixé à 4.

- **Données supplémentaires** : si vous sélectionnez cette option vous pouvez ensuite entrer le nombre de lignes et/ou de colonnes supplémentaires. Les **lignes et les colonnes supplémentaires** sont des données passives qui ne sont pas prises en compte dans les calculs de l'espace de représentation des catégories. Leurs coordonnées dans l'espace sont calculées uniquement a posteriori. Remarque : les lignes et/ou les colonnes supplémentaires doivent se trouver en fin de tableau (les dernières lignes pour les lignes supplémentaires, les dernières colonnes pour les colonnes supplémentaires).

- **Analyse d'un sous-ensemble** : si vous sélectionnez cette option vous pouvez ensuite entrer le nombre de **lignes et/ou de colonnes à exclure** pour l'analyse approfondie de certaines catégories. Voir le chapitre description pour plus de détails sur cette méthode. Remarque : les lignes et/ou les colonnes qui ne font pas partie du sous-ensemble doivent se trouver en fin de tableau (les dernières lignes pour les lignes exclues, les dernières colonnes pour les colonnes exclues).

Analyse non symétrique : cette option permet de réaliser une Analyse Non Symétrique des Correspondances, telle qu'elle a été proposée par Lauro *et al.* (1984).

- **Les lignes dépendent des colonnes** : sélectionnez cette option si vous considérez que la variable correspondant aux lignes dépend de la variable correspondant aux colonnes, et si vous voulez analyser l'association des deux variables en tenant compte de cette dépendance.

- **Les colonnes dépendent des lignes** : sélectionnez cette option si vous considérez que la variable correspondant aux colonnes dépend de la variable correspondant aux lignes, et si vous voulez analyser l'association des deux variables en tenant compte de cette dépendance.

Distance : cette option permet de calculer une Analyse Factorielle des Correspondances basée soit sur la distance du Khi², soit sur la distance de Hellinger telle que proposée par Rao (1995).

- **Khi2** : sélectionnez cette option pour réaliser une Analyse Factorielle des Correspondance classique.

- **Hellinger** : sélectionnez cette option pour réaliser une Analyse Factorielle des Correspondances basée sur la distance de Hellinger (HD). Cette option n'est pas proposée si l'option « Analyse non symétrique » a été sélectionnée ou dans le cas d'une Analyse des Correspondances Détendancée.

Pour résumer, quatre approches de l'Analyse Factorielle des Correspondances sont proposées :

- Analyse Factorielle des Correspondances classique (AFC) : ne sélectionnez pas l'option « Analyse non symétrique » et sélectionnez la distance « Khi² ».

- Analyse Non Symétrique des Correspondances (ANSC) : sélectionnez l'option « Analyse non symétrique » et sélectionnez la distance « Khi² ».
- Analyse Factorielle des Correspondances utilisant la distance de Hellinger (HD) : ne sélectionnez pas l'option « Analyse non symétrique » et sélectionnez distance de « Hellinger ».
- Analyse des Correspondances Détendancée (ACD) : sélecionnez l'option « Analyse Détendancée ».

Test d'indépendance : activez cette option si vous souhaitez que XLSTAT calcule un test d'indépendance basé sur la statistique du Khi².

- **Niveau de signification (%)** : entrez le niveau de signification pour le test (valeur par défaut : 5%).

Filtrer les facteurs : vous pouvez activer l'une des deux options suivantes afin de réduire le nombre de facteurs utilisés pour l'affichage des résultats :

- **% minimum** : activez cette option puis saisissez le pourcentage minimum que doivent représenter les facteurs retenus pour l'affichage.
- **Nombre maximum** : activez cette option pour fixer le nombre maximum de facteurs à prendre en compte pour l'affichage des résultats.

Rotation : activez cette option si vous voulez appliquer une rotation à une des matrices des coordonnées principales.

- **Nombre de facteurs** : entrez le nombre de facteurs pour lesquels la rotation sera appliquée.
- **Méthode** : choisissez la méthode de rotation à utiliser.
- **Basée sur les lignes** : activez cette option si vous voulez appliquer une rotation à partir de la matrice des coordonnées principales des lignes.
- **Basée sur les colonnes** : activez cette option si vous voulez appliquer une rotation à partir de la matrice des coordonnées principales des colonnes.

  Remarque : en choisissant la méthode « Quartimin », le choix de la matrice des coordonnées pour les calculs n'est pas possible. En effet, les calculs de la rotation s'effectuent seulement sur les coordonnées principales des lignes.
- **Normalisation de Kaiser** : activez cette option pour appliquer la normalisation de Kaiser pendant le calcul des rotations.

Onglet **Données manquantes** :

Options pour les tableaux de contingence ou croisés :

Ne pas accepter les valeurs manquantes : activez cette option pour que XLSTAT empêche la poursuite des calculs si des valeurs manquantes sont détectées.

Remplacer les valeurs manquantes par 0 : activez cette option si vous considérez que les valeurs manquantes sont équivalentes à des 0.

Remplacer les valeurs manquantes par l'espérance : activez cette option si vous souhaitez remplacer les valeurs manquantes par leur espérance. L'espérance d'une valeur manquante est donnée par :

$$E(n_{ij}) = \frac{n_{i.}n_{j.}}{n}$$

où $n_{i.}$ est la somme sur les colonnes pour la ligne i, $n_{.j}$ est la somme sur les lignes pour colonne j, et n est l'effectif total avant remplacement des valeurs manquantes.

Options pour les tableaux observations/variables :

Ne pas accepter les valeurs manquantes : activez cette option pour que XLSTAT empêche la poursuite des calculs si des valeurs manquantes sont détectées.

Supprimer les observations : activez cette option pour supprimer les observations comportant des données manquantes.

Regrouper les valeurs manquantes dans une nouvelle modalité : activez cette option pour regrouper les données manquantes dans une nouvelle modalité de la variable qualitative en question.

Onglet **Sorties** :

Options pour les tableaux observations/variables :

Statistiques descriptives : activez cette option pour afficher les statistiques descriptives pour les deux variables qualitatives sélectionnées.

Tableau disjonctif : activez cette option pour afficher le tableau disjonctif complet correspondant aux deux variables qualitatives sélectionnées.

Tri alphabétique des modalités : activez cette option pour que dans les divers résultats, les modalités soient triées alphabétiquement pour les deux variables sélectionnées.

Options communes :

Tableau de contingence : activez cette option pour afficher le tableau de contingence.

- **Vue 3D du tableau de contingence / du tableau croisé** : activez cette option pour afficher le diagramme en bâtons en 3 dimensions correspondant au tableau de contingence ou au tableau croisé.

Inertie par case : activez cette option pour afficher les inerties correspondant à chacune des cellules du tableau de contingence. Cette option n'est pas active lors de l'Analyse des Correspondances Détendancée.

Profils lignes et colonnes : activez cette option pour afficher les profils lignes et les profils colonnes. Cette option n'est pas active pour l'Analyse des Correspondances Détendancée.

Valeurs propres : activez cette option pour afficher le tableau et le graphique (*scree plot*) des valeurs propres.

Distances du Khi² (ou de Hellinger) : activez cette option pour afficher les distances du Khi² (ou de Hellinger) entre les points-lignes et entre les points-colonnes. Cette option n'est pas active pour l'Analyse des Correspondances Détendancée.

Coordonnées principales : activez cette option pour afficher les coordonnées principales des points-lignes et des points-colonnes.

Coordonnées standard : activez cette option pour afficher les coordonnées standard des points-lignes et des points-colonnes.

Contributions : activez cette option pour afficher les contributions des points-lignes et des points-colonnes aux inerties des axes factoriels. Cette option n'est pas active pour l'Analyse des Correspondances Détendancée.

Cosinus carrés : activez cette option pour afficher les cosinus carrés des points-lignes et des points-colonnes avec les axes factoriels. Cette option n'est pas active pour l'Analyse des Correspondances Détendancée.

Onglet **Graphiques** :

Sous-onglet **Cartes** :

Options pour l'analyse détendancée :

Graphiques de l'analyse détendancée : activez cette option pour afficher les graphiques lors d'une Analyse des Correspondances Détendancée.

- **Lignes et colonnes** : activez cette option pour afficher un graphique sur lequel sont affichés les points-lignes et les points-colonnes.
- **Lignes** : activez cette option pour afficher un graphique sur lequel sont représentés uniquement les points-lignes.
- **Colonnes** : activez cette option pour afficher un graphique sur lequel sont représentés uniquement les points-colonnes.

Etiquettes : activez cette option pour afficher les libellés des modalités sur les graphiques.

- **Etiquettes colorées** : activez cette option pour que les étiquettes soient de la même couleur que les points correspondants.

Options pour les autres analyses :

Graphiques symétriques : activez cette option pour afficher les graphiques pour lesquels les points-lignes et les points-colonnes jouent un rôle symétrique. Ces graphiques sont aussi appelés graphiques barycentriques. Ces graphiques utilisent les coordonnées principales des points-lignes et des points-colonnes.

- **Lignes et colonnes** : activez cette option pour afficher un graphique sur lequel sont représentés les points-lignes et les points-colonnes.

- **Lignes** : activez cette option pour afficher un graphique sur lequel sont représentés uniquement les points-lignes.

- **Colonnes** : activez cette option pour afficher un graphique sur lequel sont représentés uniquement les points-colonnes.

Graphiques asymétriques : activez cette option pour afficher les graphiques pour lesquels les points-lignes et les points-colonnes jouent un rôle asymétrique. Ces graphiques sont aussi appelés graphiques pseudo-barycentriques. Ces graphiques utilisent les coordonnées principales d'une part et les coordonnées standard d'autre part.

- **Lignes** : activez cette option pour afficher un graphique sur lequel sont affichés les points-lignes avec leurs coordonnées principales, et les points-colonnes avec leurs coordonnées standard.

- **Colonnes** : activez cette option pour afficher un graphique sur lequel sont affichés les points-colonnes avec leurs coordonnées principales, et les points-lignes avec leurs coordonnées standard.

- **Vecteurs** : activez cette option pour afficher des vecteurs pour les coordonnées standard sur les graphiques asymétriques.

- **Facteur d'allongement** : activez cette option pour jouer sur la longueur des vecteurs affichés.

Biplot de contribution : activez cette option pour afficher les biplots de contribution. Ces graphiques utilisent les coordonnées principales d'une part et les coordonnées de contribution d'autre part. Les coordonnées de contribution correspondent aux coordonnées standard multipliées par la racine carrée du poids relatif de la modalité.

- **Lignes** : activez cette option pour afficher un graphique sur lequel sont affichés les points-lignes avec leurs coordonnées principales, et les points-colonnes avec leurs coordonnées de contribution.

- **Colonnes** : activez cette option pour afficher un graphique sur lequel sont affichés les points-colonnes avec leurs coordonnées principales, et les points-lignes avec leurs coordonnées de contribution.

Ellipses de confiance : activez cette option pour afficher les ellipses de confiance afin d'identifier les catégories qui contribuent à la dépendance entre les variables.

- **Lignes** : activez cette option pour afficher les ellipses de confiance pour les points-lignes sur le graphique symétrique « Lignes ».

- **Colonnes** : activez cette option pour afficher les ellipses de confiance pour les points-colonnes sur le graphique symétrique « Colonnes ».

Graphiques de rotation : cette option n'est pas active quand la méthode « Quartimin » est sélectionnée. Activez cette option pour afficher les graphiques : le biplot regroupant les coordonnées des points-lignes et les coordonnées des points-colonnes calculées après rotation, le graphique sur lequel sont uniquement affichés les points-lignes, et le graphique sur lequel sont uniquement affichés les points-colonnes. Dans le cas où l'option « Basée sur les

lignes » est sélectionnée, les coordonnées des lignes affichées sont les coordonnées principales et les coordonnées des colonnes sont les coordonnées standard.

Sous-onglet **Options des lignes** :

Ces options ne sont pas actives pour le cas de l'Analyse des Correspondances Détendancée.

Filtrer les lignes : activez cette option pour fixer le nombre de lignes affichées :

Options pour les tableaux de contingence ou croisés :

- **Aléatoire** : les lignes à afficher sont sélectionnées de manière aléatoire. Le « Nombre de lignes » doit alors être saisi.
- **N premières lignes** : les N premières lignes du tableau de contingence sont affichées. Le « Nombre de lignes » N doit alors être saisi.
- **N dernières lignes** : les N dernières lignes du tableau de contingence sont affichées. Le « Nombre de lignes » N doit alors être saisi.
- **Variable de groupe** : si vous choisissez cette option, vous devez ensuite sélectionner une variable indicatrice composée de 1 pour les lignes à afficher, et de 0 pour les lignes à ne pas afficher. Cette variable doit comporter le même nombre de lignes que le tableau de contingence.

Options communes :

- **Somme(Cos2)>** : choisissez cette option pour que seules les lignes ayant une somme des cosinus carrés supérieure à une valeur à saisir entre 0 et 1 soient affichées sur les graphiques.

Redimensionner les points lignes avec Cos2 : activez cette option pour redimensionner les points-lignes. La taille des points-lignes sera proportionnelle à la somme des cosinus carrés sur les dimensions représentées.

Libellés des lignes : activez cette option pour afficher les libellés des lignes sur les graphiques.

- **Etiquettes colorées** : activez cette option pour que les étiquettes des lignes soient de la même couleur que les points correspondants.

Sous-onglet **Options des colonnes** :

Ces options ne sont pas actives pour le cas de l'Analyse des Correspondances Détendancée.

Filtrer les colonnes : activez cette option pour fixer le nombre de colonnes affichées :

Options pour les tableaux de contingence ou croisés :

- **Aléatoire** : les colonnes à afficher sont sélectionnées de manière aléatoire. Le « Nombre de colonnes » doit alors être saisi.
- **N premières colonnes** : les N premières colonnes du tableau de contingence sont affichées. Le « Nombre de colonnes » N doit alors être saisi.

- **N dernières colonnes** : les N dernières colonnes du tableau de contingence sont affichées. Le « Nombre de colonnes » N doit alors être saisi.

- **Variable de groupe** : si vous choisissez cette option, vous devez ensuite sélectionner une variable indicatrice composée de 1 pour les colonnes à afficher, et de 0 pour les colonnes à ne pas afficher. Cette variable doit comporter le même nombre de lignes que le tableau de contingence n'a de colonnes.

Options communes :

- **Somme(Cos2)>** : choisissez cette option pour que seules les colonnes ayant une somme des cosinus carrés supérieure à une valeur à saisir entre 0 et 1 soient affichées sur les graphiques.

Redimensionner les points colonnes avec Cos2 : activez cette option pour redimensionner les points-colonnes. La taille des points-colonnes est proportionnelle à la somme des cosinus carrés sur les dimensions représentées.

Libellés des colonnes : activez cette option pour afficher les libellés des colonnes sur les graphiques.

- **Etiquettes colorées** : activez cette option pour que les étiquettes des colonnes soient de la même couleur que les points correspondants.

Résultats

Statistiques descriptives : ce tableau n'est affiché que si les données d'entrée correspondent à un tableau observations/variables. Sont affichés le nom des modalités de chaque variable, la fréquence de chaque modalité et le pourcentage que représente chaque modalité pour la variable.

Tableau disjonctif : ce tableau n'est affiché que si les données d'entrée correspondent à un tableau observations/variables. Ce tableau est un tableau intermédiaire permettant d'aboutir au tableau de contingence des deux variables sélectionnées.

Tableau de contingence : le tableau de contingence est affiché. Le diagramme en bâtons en **3 dimensions** en est la représentation graphique.

Inertie par case : le tableau des inerties par case est affiché. La somme des inerties est égale à la statistique du Khi² divisée par la fréquence totale (somme des cellules du tableau de contingence).

Test d'indépendance entre les lignes et les colonnes : ce test permet de déterminer, sur la base de la statistique du Khi², si l'on doit rejeter l'hypothèse nulle selon laquelle les lignes et les colonnes du tableau sont indépendantes. Une interprétation détaillée est fournie automatiquement.

Valeurs propres et pourcentages d'inertie : les valeurs propres et le graphique (scree plot) correspondant sont affichés. Seules les valeurs propres non triviales sont affichées. Si un filtrage a été demandé, il est appliqué aux résultats qui suivent. Dans le cas de l'Analyse des Correspondances Détendancée, seules les valeurs propres sont affichées.

Remarque : la somme des valeurs propres affichées est égale à l'inertie totale. A chaque valeur propre correspond un axe principal représentant un pourcentage donné de l'inertie totale. On peut ainsi mesurer le pourcentage cumulé d'inertie totale correspondant à un nombre croissant de dimensions.

Une série de résultats est ensuite affichée, d'abord pour les points-lignes, puis pour les points-colonnes :

Poids, distances et distances quadratiques à l'origine, inerties et inerties relatives : ce tableau contient des statistiques de base pour les points-lignes (puis pour les points-colonnes).

Profils : dans ce tableau sont affichés les profils. La moyenne des profils est également affichée sauf dans lors d'une Analyse Factorielle des Correspondances basée sur la distance de Hellinger.

Distances du Khi² (ou de Hellinger) : dans ce tableau sont affichées les distances du Khi² (ou de Hellinger) entre les profils.

Coordonnées principales : dans ce tableau sont affichées les coordonnées principales. Ces coordonnées sont utilisées pour la création des graphiques symétriques (ou barycentriques) et asymétriques (ou pseudo-barycentriques).

A la fin du tableau des coordonnées principales des lignes (ou des colonnes), vous trouverez le bouton suivant :  .

Ce bouton vous permet d'ouvrir automatiquement la boite de dialogue pré-remplie de la CAH (Classification Ascendante Hiérarchique) afin d'effectuer une classification sur les coordonnées factorielles des lignes (ou des colonnes).

Coordonnées standard : ces coordonnées correspondent aux précédentes à un facteur près. Le facteur est la racine carrée de l'inverse de la valeur propre correspondante. Ces coordonnées sont utilisées pour la création des graphiques asymétriques (ou pseudo-barycentriques).

Contributions : les contributions sont une aide à l'interprétation. Les modalités ayant influencé le plus la construction des axes sont celles dont les contributions sont les plus élevées. On pourra se contenter d'interpréter les résultats des modalités pour lesquelles les contributions sont supérieures aux poids relatifs affichés dans la première colonne.

Cosinus carrés : comme pour les autres méthodes factorielles, l'analyse des cosinus carrés permet d'éviter des erreurs d'interprétation dues à des effets de projection. Si les cosinus carrés associés aux axes utilisés sur un graphique sont faibles, on évitera d'interpréter la position du point-ligne ou du point-colonne correspondant.

Les graphiques constituent le but ultime de l'Analyse Factorielle des Correspondances, car ils permettent d'accélérer considérablement l'interprétation des résultats.

Graphiques symétriques : aussi appelés représentations barycentriques, ces graphiques utilisent exclusivement les coordonnées principales. En fonction des choix effectués dans la boîte de dialogue, sont affichés, un graphique symétrique mélangeant points-lignes et points-

colonnes, un graphique des points-lignes, et un graphique des points-colonnes. Le pourcentage d'inertie correspondant à chacun des axes concernés et le pourcentage d'inertie cumulée du graphique sont affichés. La proximité entre deux modalités sur le graphique est représentative de leur association. Si l'option « Ellipses de confiance » a été sélectionnée, des ellipses de confiances sont dessinée autour des points sur les graphiques représentant les points-lignes et les points-colonnes seuls. Les ellipses de confiance permettent l'identification des catégories contribuant à la structure d'association entre les variables. Les ellipses reflètent l'information contenue dans les dimensions non représentées sur la carte.

Graphiques asymétriques : aussi appelés représentations pseudo-barycentriques, ces graphiques utilisent d'une part les coordonnées principales (pour les points-lignes ou les points-colonnes) et d'autre part les coordonnées standard (respectivement pour les points-colonnes ou les points-lignes). Le pourcentage d'inertie correspondant à chacun des axes concernés et le pourcentage d'inertie cumulée du graphique sont affichés. Le nom du graphique, par exemple « Graphique asymétrique des lignes » indique les points qui font l'objet d'une interprétation : sur un « Graphique asymétrique des lignes », on étudiera la façon dont les points lignes sont positionnés par rapport aux vecteurs des modalités, ces derniers donnant des directions. Si deux points-lignes sont dans la direction d'un vecteur modalité, la modalité correspondant au point-ligne qui est le plus éloigné de l'origine est celle qui est la plus liée à la modalité correspondant au vecteur.

Biplots de contribution : ces biplots, mis au point par Greenacre, permettent d'éviter certains problèmes des graphiques asymétriques, tout en faisant ressortir les points contribuant le plus à la construction de l'axe (les points-colonnes dans le cas d'un biplot de contribution sur les lignes et vice-versa).

Remarque : Pour une Analyse des Correspondances Détendancée, les graphiques affichés ne sont pas les mêmes. L'option « Lignes et colonnes » permet d'afficher simultanément les coordonnées principales des lignes et les coordonnées standard des colonnes. L'option « Lignes » permet d'afficher seulement les coordonnées principales des lignes et, respectivement, l'option « Colonnes » permet d'afficher les coordonnées standard des colonnes. De plus, le pourcentage d'inertie correspondant à chacun des axes concernés et le pourcentage d'inertie cumulée du graphique ne sont pas affichés. En effet, les étapes de détendance provoquent une distorsion de la variance ce qui empêche l'interprétation des valeurs propres comme une partition de celle-ci. Les valeurs propres doivent donc être interprétées comme indiquant l'importance relative de chaque axe.

Dans le cas où une rotation a été demandée, les résultats de la rotation sont affichés, avec en premier la **matrice de rotation** appliquée aux coordonnées des lignes et des colonnes. Suivent ensuite les pourcentages modifiés de variabilité associés à chacun des axes concernés par la rotation. Dans les tableaux suivants sont affichées les coordonnées, les contributions et les cosinus des lignes et des colonnes après rotation. Si l'option « Graphiques de rotation » a été sélectionnée, les graphiques représentant les coordonnées des lignes et des colonnes sont affichés.

Exemple

Un exemple d'utilisation de l'Analyse Factorielle des Correspondances est disponible sur le Centre d'aide XLSTAT à l'adresse suivante :

http://www.xlstat.com/demo-caf.htm

Bibliographie

Balbi S. (1997). Graphical Displays in Non Symmetric Correspondence Analysis. In: Blasius J. and Greenacre M. (eds.), Visualisation of Categorical Data. Academic Press, San Diego. 297-309.

Beh E. J. & Lombardo R. (2015). Confidence Regions and Approximate p-values for Classical and Non Symmetric Correspondence Analysis.*Communications in Statistics-Theory and Methods*, ***44*** **(1)**, 95-114.

Benzécri J.P. (1969). Statistical Analysis as a Tool to Make Patterns Emerge from Data. In Watanabe S. (ed.), Methodologies of Pattern Recognition. Academic Press, New York. 35-60.

Benzécri J.P. (1973). L'Analyse des Données, Tome2: L'analyse des correspondances. Dunod, Paris.

Benzécri J.P. (1992). Correspondence Analysis Handbook. Marcel Decker, New York.

Cuadras C. M. & Cuadras i Pallejà D. (2008). A unified approach for representing rows and columns in contingency tables.

Goodman, L. A. and Kruskal, W. H. (1954). Measures of association for cross classifications. *Journal of the American Statistical Association.* **49**, 732-764.

Greenacre M. J. (1984). Theory and Applications of Correspondence Analysis. Academic Press, London.

Greenacre M. J. (1993). Correspondence Analysis in Practice. Academic Press, London.

Greenacre M. J. and Pardo R. (2006). Subset correspondence analysis: Visualizing relationships among a selected set of response categories from a questionnaire survey. *Sociological Methods & Research*, **35** (2), 193-218.

Hill M. O. and Gauch Jr. H. G. (1980). Detrended correspondence analysis: An improved ordination technique. *Vegetation*., **42**, 47-58

Lauro C., Balbi S. (1999). The analysis of structured qualitative data. *Applied Stochastic Models and Data Analysis*. **15**, 1-27.

Lauro N.C., D'Ambra L. (1984). Non-symmetrical Correspondence Analysis. In: Diday E. *et al.* (eds.), Data Analysis and Informatics, **III**, North Holland, Amsterdam. 433-446.

Lebart L., Morineau A. and Piron M. (1997). Statistique Exploratoire Multidimensionnelle, 2ème édition. Dunod, Paris. 67-107.

Lorenzo-Seva U., Van de Velden M., Kiers H. A. L. (2009). Oblique rotation in correspondence analysis: A step forward in the search for the simplest interpretation. *British Journal of Mathematical and Statistical Psychology*, **62**, 583-600.

Rao, C. R. (1995). A review of canonical coordinates and an alternative to correspondence analysis using Hellinger distance. *Questiió: Quaderns d'Estadística, Sistemes, Informatica i Investigació Operativa*, **19** (1), 23-63.

Saporta G. (1990). Probabilités, Analyse des Données et Statistique. Technip, Paris. 199-216.

Van de Velden M. and Kiers H. A. L. (2005). Rotation in correspondence analysis. *Journal of Classification*, **22**, 251-271.

Analyse des Correspondances Multiples (ACM)

Utilisez ce module pour représenter graphiquement l'association entre les modalités (aussi appelées catégories) d'au moins deux variables qualitatives. L'ACM peut aussi être utilisée pour transformer des données qualitatives en des données quantitatives utilisables ensuite par des méthodes de classification.

Dans cette section :

Description

L'Analyse des Correspondances Multiples (ACM) est une méthode qui permet d'étudier l'association entre au moins deux variables qualitatives. L'ACM est aux variables qualitatives ce que l'Analyse en Composantes Principales est aux variables quantitatives. Elle permet en effet d'aboutir à des cartes de représentation sur lesquelles ont peut visuellement observer les proximités entre les catégories des variables qualitatives et les observations.

L'Analyse des Correspondances Multiples (ACM) peut aussi être vue comme la généralisation de l'AFC au cas où l'on a plus de deux variables. S'il est possible de synthétiser un tableau à n individus et p ($p > 2$) variables qualitatives dans un tableau dont la structure est proche d'un tableau de contingence, il est beaucoup plus commun en ACM de partir d'un tableau observations/variables (par exemple à la suite d'une enquête, où l'on a posé p questions à n individus). XLSTAT permet aussi de travailler à partir d'un tableau disjonctif complet.

La construction du tableau disjonctif complet est de toute manière l'une des étapes préalables au calcul de l'ACM. Les p variables qualitatives sont éclatées en p tableaux disjonctifs $Z_1, Z_2, \ldots, Zp$, composés d'autant de colonnes qu'il y a de modalités pour chacune des variables. A chaque fois qu'une modalité m de la j-ème variable correspond à un individu i, on affecte 1 à $Z_j(i, m)$. Les autres valeurs de Z_j sont nulles. Les p tableaux disjonctifs sont alors concaténés en un tableau disjonctif complet.

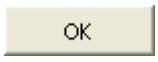
A partir du tableau disjonctif complet sont calculées les coordonnées des modalités des variables qualitatives, ainsi que les coordonnées des observations dans un espace de représentation optimal pour le critère d'inertie. Dans le cas de l'ACM on montre que l'inertie est égale au nombre moyen de modalités moins un. Elle ne dépend donc pas uniquement de l'association entre les variables. XLSTAT propose également de réaliser l'ACM en se basant sur

la matrice de Burt au lieu d'utiliser le tableau disjonctif complet. Greenacre (1993) a proposé une mesure ajustée de l'inertie, inspirée de la Joint Correspondence Analysis (JCA). Cet ajustement permet d'avoir des pourcentages plus élevés et plus informatifs pour les axes de représentation.

L'**analyse d'un sous-ensemble** de modalités (ou catégories), est une méthode très récemment mise au point par Greenacre et Pardo (2006), qui permet de focaliser l'étude sur quelques catégories uniquement, tout en prenant en compte toutes les données du tableau de données initial. XLSTAT vous permet de sélectionner les catégories sur lesquelles vous souhaitez focaliser l'analyse.

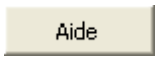
Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.



: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.

Onglet **Général**:

Le champ principal de saisie des données vous permet de sélectionner alternativement deux types de tableaux :

Tableau observations/variables : choisissez cette option si vos données correspondent à un tableau comprenant n observations décrites par p variables qualitatives. Si des en-têtes de colonnes ont été sélectionnés, veuillez vérifier que l'option « Libellés des variables » est activée.

Tableau disjonctif : choisissez cette option si vos données correspondent à un tableau disjonctif. Si des en-têtes de colonnes ont été sélectionnés, veuillez vérifier que l'option « Libellés des variables » est activée. Si cette option est activée, les observations

supplémentaires et les variables qualitatives supplémentaires devront également être sous la forme d'un tableau disjonctif.

Poids : activez cette option si vous voulez pondérer les observations. Si vous n'activez pas cette option, les poids seront tous considérés comme valant 1. Les poids doivent être impérativement supérieurs ou égaux à 0. Si un en- tête de colonne a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des variables : activez cette option si vous avez inclus les libellés des variables (cas d'un tableau observations/variables) ou les libellés des modalités (cas d'un tableau disjonctif) dans la sélection.

Libellés des observations : activez cette option si vous voulez utiliser des libellés d'observations disponibles pour l'affichage des résultats. Si l'option « Libellés des variables » est activée, la première cellule de la sélection doit comprendre un en-tête. Si vous n'activez pas cette option, des libellés seront automatiquement créés (Obs1, Obs2, …).

Onglet **Options**:

Type d'ACM :

- **Tableau disjonctif** : si vous sélectionnez cette option, l'ACM sera réalisée à l'aide du tableau disjonctif complet du tableau observations/variables.
- **Tableau de Burt** : si vous sélectionnez cette option, l'ACM sera réalisée à l'aide du tableau de Burt du tableau observations/variables.
- **Inertie ajustée** : si vous sélectionnez cette option, l'ACM sera réalisée à l'aide du tableau de Burt et les inerties obtenues seront ajustées à l'aide de la méthode proposée par Greenacre (1993).

Analyse approfondie :

- **Aucune** : si vous sélectionnez cette option, l'ACM sera réalisée sur toutes les modalités des variables qualitatives actives.
- **Analyse d'un sous-ensemble** : si vous sélectionnez cette option, au cours des calculs XLSTAT vous demandera de préciser quelles sont les modalités (ou catégories)

constitutives du sous-ensemble à analyser.

Tri alphabétique des modalités : activez cette option pour que dans les divers résultats, les modalités soient triées alphabétiquement pour chacune des variables.

Libellés Variable-Modalité : activez cette option pour utiliser des libellés longs pour l'affichage des résultats. Les libellés Variable-Modalité sont composés du nom de la variable comme préfixe, et de la modalité comme suffixe.

Filtrer les facteurs : vous pouvez activer l'une ou les deux options suivantes afin de réduire le nombre de facteurs utilisés pour l'affichage des résultats :

- **% minimum** : activez cette option puis saisissez le pourcentage minimum que doivent représenter les facteurs retenus pour l'affichage.
- **Nombre maximum** : activez cette option pour fixer le nombre maximum de facteurs à prendre en compte pour l'affichage des résultats.
- **1/p** : activez cette option pour ne prendre en compte que les facteurs dont la valeur propre correspondante est supérieure à 1/p, où p est le nombre de variables actives. Cette option est activée par défaut.

Onglet **Données supplémentaires** :

Observations supplémentaires : activez cette option si vous voulez représenter des individus supplémentaires par le calcul de leurs coordonnées. Ces individus ne sont pas pris en compte pour le calcul des axes factoriels (observations passives, par opposition à observations actives). Si des libellés de variables sont présents pour les observations supplémentaires vous devez activer l'option « Libellés des variables pour les obs. ». Vous pouvez également choisir des libellés des observations supplémentaires pour l'affichage des résultats.

Variables supplémentaires : activez cette option si vous voulez calculer les coordonnées a posteriori pour des variables qui ne sont pas prises en compte pour le calcul des axes factoriels (variables passives, par opposition aux variables actives).

- **Quantitatives** : activez cette option si vous disposez de variables quantitatives supplémentaires. Si des en-têtes de colonnes ont été sélectionnés pour le tableau principal, veillez à ce qu'un libellé soit aussi présent pour les variables de cette sélection.
- **Qualitatives** : activez cette option si vous disposez de variables qualitatives supplémentaires. Si des en-têtes de colonnes ont été sélectionnés pour le tableau principal, veillez à ce qu'un libellé soit aussi présent pour les variables de cette sélection.

Onglet **Données manquantes** :

Ne pas accepter les valeurs manquantes : activez cette option pour que XLSTAT empêche la poursuite des calculs si des valeurs manquantes sont détectées.

Supprimer les observations : activez cette option pour supprimer les observations comportant des données manquantes.

Remplacer les valeurs manquantes : activez cette option pour remplacer les valeurs manquantes. Pour les variables quantitatives supplémentaires, les données manquantes sont remplacées par la moyenne de la variable quantitative concernée, tandis que pour les variables qualitatives du tableau initial (variables actives) ou pour les variables qualitatives supplémentaires (variables passives), une nouvelle catégorie « Manquant » est créée pour les variables qualitatives en question.

Onglet **Sorties**:

Statistiques descriptives : activez cette option pour afficher les statistiques descriptives pour les variables sélectionnées.

Tableau disjonctif : activez cette option pour afficher le tableau disjonctif complet correspondant aux deux variables qualitatives sélectionnées.

Tableau de Burt : activez cette option pour afficher le tableau de Burt.

Vue 3D du tableau de Burt : activez cette option pour afficher le diagramme en bâton en 3 dimensions correspondant au tableau de Burt. Ces graphiques utilisent les coordonnées principales.

Valeurs propres : activez cette option pour afficher le tableau et le graphique (*scree plot*) des valeurs propres.

Affichez les résultats pour :

- **Observations et variables** : activez cette option pour afficher les résultats concernant les observations et les variables.
- **Observations** : activez cette option pour afficher uniquement les résultats concernant les observations.
- **Variables** : activez cette option pour afficher uniquement les résultats concernant les variables.

Coordonnées principales : activez cette option pour afficher les coordonnées principales.

Coordonnées standard : activez cette option pour afficher les coordonnées standard.

Contributions : activez cette option pour afficher les contributions.

Cosinus carrés : activez cette option pour afficher les cosinus carrés.

Valeurs test : activez cette option pour afficher les valeurs test pour les variables.

- **Niveau de signification (%)** : entrez le niveau de signification pour déterminer si les valeurs test sont significatives ou non.

Onglet **Graphiques** :

Sous-Onglet **Variables** :

Carte factorielle des modalités : activez cette option pour afficher le graphique des coordonnées principales des modalités des variables qualitatives actives et supplémentaires.

- **Etiquettes** : activez cette option pour que les étiquettes des noms des modalités soient affichées sur le graphique.
- **Etiquettes** : activez cette option pour que les étiquettes des noms des modalités soient de la même couleur que les points correspondants.
- **Colorer par groupe** : activez cette option si vous souhaitez colorer les variables en fonction d'une variable de groupe, sélectionnez un vecteur colonne de taille égale au nombre de variables actives. Si des en-têtes de colonnes ont été sélectionnés pour le tableau principal, veillez à ce qu'un libellé soit aussi présent pour la variable de cette sélection.
- **Taille des points = f(Cos2)** : activez cette option si vous souhaitez que la taille des points correspondants aux variables soit proportionnelle à leur cosinus carré sur le plan sélectionné.
- **Relier les modalités** : activez cette option pour que les modalités de chaque variable soient reliées entre elles. Cette option permet de repérer plus rapidement les modalités appartenant à une variable.

Cercle des corrélations des var. supp : activez cette option pour représenter le cercle des corrélations des variables supplémentaires quantitatives.

Filtrer : activez cette option pour filtrer les variables affichées :

- **Aléatoire** : les variables à afficher sont sélectionnées de manière aléatoire. Le « Nombre de variables » N doit alors être saisi.
- **N premières variables** : les N premières variables sont affichées. Le « Nombre de variables » N doit alors être saisi.
- **N dernières variables** : les N dernières variables sont affichées. Le « Nombre de variables » N doit alors être saisi.
- **Variable de groupe** : si vous choisissez cette option, vous devez ensuite sélectionner une variable indicatrice composée de 1 pour les variables à afficher, et de 0 pour les

variables à ne pas afficher.

Sous-Onglet **Observations** :

Carte factorielle des observations : activez cette option pour afficher le graphique des coordonnées principales des observations actives et des observations supplémentaires.

- **Etiquettes** : activez cette option pour afficher les étiquettes des observations à côté des points.
- **Etiquettes colorées** : activez cette option pour que les étiquettes soient de la même couleur que les points correspondants
- **Colorer par groupe** : activez cette option si vous souhaitez colorer les observations en fonction d'une variable de groupe, sélectionnez un vecteur colonne de taille égale au nombre d'observations actives. Si des en-têtes de colonnes ont été sélectionnés pour le tableau principal, veillez à ce qu'un libellé soit aussi présent pour la variable de cette sélection.
- **Taille des points = f(Cos2)** : activez cette option si vous souhaitez que la taille des points correspondants aux observations soit proportionnelle à leur cosinus carré sur le plan sélectionné.

Filtrer : activez cette option pour filtrer les observations affichées :

- **Aléatoire** : les variables à afficher sont sélectionnées de manière aléatoire. Le « Nombre de observations » N doit alors être saisi.
- **N premières lignes** : les N premières variables sont affichées. Le « Nombre de observations » N doit alors être saisi.
- **N dernières lignes** : les N dernières variables sont affichées. Le « Nombre de observations » N doit alors être saisi.
- **Variable de groupe** : si vous choisissez cette option, vous devez ensuite sélectionner une variable indicatrice composée de 1 pour les observations à afficher, et de 0 pour les observations à ne pas afficher.
- **Somme(Cos2)>** : choisissez cette option pour que, seules les observations ayant une somme des cosinus carrés supérieure à une valeur à saisir entre 0 et 1, soient affichées sur les graphiques de représentation des observations.

Sous-Onglet **Biplots** :

Biplots : activez cette option pour afficher les graphiques de représentation simultanée des observations et des modalités.

- **Graphique symétrique** : choisissez cette option pour représenter sur un même graphique les coordonnées principales des modalités et les coordonnées principales des

observations sur le même graphique.

- **Graphiques asymétriques** : activez cette option pour afficher les graphiques asymétriques. Ces graphiques utilisent les coordonnées principales et les coordonnées standardisées. Deux types de représentation sont disponibles :

- **Graphique asymétriques des lignes** : activez cette option pour afficher un graphique sur lequel sont affichées les observations avec leurs coordonnées principales, et les modalités des variables qualitatives avec leurs coordonnées standardisées.

- **Graphique asymétrique des colonnes** : activez cette option pour afficher un graphique sur lequel sont affichées les modalités des variables qualitatives avec leurs coordonnées principales, et les observations avec leurs coordonnées standardisées.

Options pour les variables :

- **Libellés des modalités** : activez cette option pour afficher les étiquettes avec les noms des modalités.

- **Vecteurs** : activez cette option pour afficher des vecteurs reliant le centre du graphique et les points correspondants aux coordonnées des modalités.

- **Variables supplémentaires** : si vous avez inclus des variables supplémentaires dans l'analyse, activez cette option pour les afficher sur les biplots.

- **Filtrer les variables** : si vous avez utilisé une variable de filtrage pour les variables, cette même variable sera utilisée pour filtrer les variables sur les biplots.

Options pour les observations :

- **Libellés des observations** : activez cette option pour afficher les étiquettes avec les noms des observations.

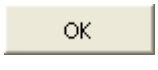
- **Vecteurs** : activez cette option pour afficher des vecteurs reliant le centre du graphique et les points correspondants aux coordonnées des observations.

- **Observations supplémentaires** : si vous avez inclus des observations supplémentaires dans l'analyse, activez cette option pour les afficher sur les biplots.

- **Filtrer les observations** : si vous avez utilisé une variable de filtrage pour les observations, cette même variable sera utilisée pour filtrer les observations sur les biplots.

Boîte de dialogue (sous-ensemble de modalités)

Cette boîte de dialogue est affichée si vous avez sélectionné l'option **Analyse approfondie** / **Analyse d'un sous-ensemble** dans la boîte de dialogue principale.



: cliquez sur ce bouton pour reprendre les calculs.

Aide : cliquez sur ce bouton pour afficher l'aide.

La **liste des modalités** correspondant à l'ensemble des variables qualitatives actives est affichée. Sélectionnez alors les modalités sur lesquelles vous voulez que l'analyse soit focalisée.

Toutes : cliquez sur ce bouton pour sélectionner toutes les catégories.

Aucune : cliquez sur ce bouton pour désélectionner toutes les catégories.

Résultats

Statistiques descriptives : ce tableau n'est affiché que si les données d'entrée correspondent à un tableau observations/variables.

Tableau disjonctif : ce tableau n'est affiché que si les données d'entrée correspondent à un tableau observations/variables. Ce tableau est un tableau intermédiaire permettant d'aboutir au tableau de contingence des deux variables sélectionnées.

Tableau de Burt : le tableau de Burt est affiché si l'option correspondante a été activée. Le diagramme en bâtons en **3 dimensions** en est la représentation graphique.

Valeurs propres et pourcentages d'inertie : les valeurs propres, les pourcentages d'inertie et le graphique correspondant (scree plot) sont affichés. Seules les valeurs propres non triviales sont affichées. Si un filtrage a été demandé il est appliqué aux résultats qui suivent.

Une série de résultats est ensuite affichée, d'abord pour les variables, puis pour les observations :

Coordonnées principales : dans ce tableau sont affichées les coordonnées principales. Ces coordonnées sont utilisées pour la création des graphiques symétriques (ou barycentriques) et asymétriques (ou pseudo-barycentriques) où elles représentent les projections des profils.

A la fin du tableau des coordonnées des observations, vous trouverez le bouton suivant :  .

Ce bouton vous permet d'ouvrir automatiquement la boite de dialogue pré-remplie de la CAH (Classification Ascendante Hiérarchique) afin d'effectuer une classification sur les coordonnées factorielles des observations.

Coordonnées standard : ces coordonnées correspondent aux précédentes à un facteur près. Le facteur est la racine carrée de l'inverse de la valeur propre correspondante. Ces coordonnées sont utilisées pour la création des graphiques asymétriques (ou pseudo-barycentriques) où elles représentent les projections des profils normés.

Contributions : les contributions sont une aide à l'interprétation. Les modalités ayant influencé le plus la construction des axes sont celles dont les contributions sont les plus élevées. On pourra se contenter d'analyser les contributions qui sont supérieures aux poids relatifs affichés dans la seconde colonne.

Indice d'homogénéité des axes: : Cet indice développé par nos équipes est très utiles pour déterminer si les contributions des observations sont homogènes pour les différents axes. Il est construit comme la proportion d'observations ayant une contribution absolue > 1/n. Un indice au dessus de 0.4 indique une très bonne homogénéité avec des observations bien représentées. En revanche, un indice inférieur à 0.1 doit être une alerte pour l'utilisateur qui devrait vérifier si il n'a pas de valeurs extrêmes sur les variables construisant l'axe qui fausseraient son interprétation (les valeurs extrêmes seraient alors les observations se démarquant des autres sur l'axe en question).

Cosinus carrés : comme pour les autres méthodes factorielles, l'analyse des cosinus carrés permet d'éviter des erreurs d'interprétation dues à des effets de projection. Si les cosinus carrés associés aux axes utilisés sur un graphique sont faibles, on évitera d'interpréter la position de l'observation ou de la variable en question.

Les graphiques constituent le but ultime de l'Analyse des Correspondances Multiples, car ils permettent d'accélérer considérablement l'interprétation des données.

Graphiques symétriques : aussi appelés représentations barycentriques, ces graphiques utilisent exclusivement les coordonnées principales. En fonction des choix effectués dans la boîte de dialogue, sont affichés, un graphique symétrique mélangeant observations et variables, un graphique des observations, et un graphique des variables. Le pourcentage d'inertie correspondant à chacun des axes concernés et le pourcentage d'inertie cumulée du graphique sont affichés.

Graphiques asymétriques : aussi appelés représentations pseudo- barycentriques, ces graphiques utilisent d'une part les coordonnées principales pour les observations et d'autre part les coordonnées standard pour les variables, et réciproquement. Le pourcentage d'inertie correspondant à chacun des axes concernés et le pourcentage d'inertie cumulée du graphique sont affichés. Sur un « graphique asymétrique des observations », on étudiera la façon dont les observations sont positionnées par rapport aux vecteurs des modalités, ces derniers indiquant des directions. Si deux observations sont dans la direction d'un vecteur modalité, l'observation qui est la plus éloignée de l'origine est celle pour laquelle la modalité a le plus vraisemblablement été choisie.

Exemple

Un exemple d'utilisation de l'Analyse des Correspondances Multiples est disponible sur le Centre d'aide XLSTAT à l'adresse suivante :

http://www.xlstat.com/demo-mcaf.htm

Bibliographie

Greenacre M. J. (1984). Theory and Applications of Correspondence Analysis. Academic Press, London.

Greenacre M. J. (1993). Correspondence Analysis in Practice. Academic Press, London.

Greenacre, M.J. (1993). Multivariate Generalizations of Correspondence Analysis. In: Multivariate Analysis: Future Directions 2 (Eds: C.M. Cuadras and C.R. Rao), Elsevier Science, Amsterdam. 327-340.

Greenacre M. J. and Pardo R. (2006). Multiple correspondence analysis of subsets of response categories. In Multiple Correspondence Analysis and Related Methods (eds Michael Greenacre and Jörg Blasius), Chapman & Hall/CRC, London, 197-217.

Lebart L., Morineau A. and Piron M. (1997). Statistique Exploratoire Multidimensionnelle, 2ème édition. Dunod, Paris. 108-145.

Saporta G. (1990). Probabilités, Analyse des Données et Statistique. Technip, Paris. 217-239.

Multidimensional Scaling (MDS)

Utilisez le Multidimensional Scaling (MDS) pour représenter dans un espace à deux ou trois dimensions des objets pour lesquels seule une matrice de proximité (similarité ou dissimilarité) est disponible.

Dans cette section :

Description

Le Multidimensional Scaling (MDS) permet de passer d'une matrice de **proximité** (similarité ou dissimilarité) entre une série de N objets aux coordonnées de ces mêmes objets dans un espace à p dimensions. On fixera en général p à 2 ou 3 afin de pouvoir facilement visualiser les objets. Par exemple, avec le MDS, il est possible de reconstituer très précisément la position de villes sur une carte à partir des distances kilométriques (la dissimilarité est alors une distance euclidienne) entre les villes, à une rotation et une symétrie près.

L'exemple ci-dessus a pour seul intérêt de montrer la performance de la méthode, et de faire comprendre son esprit. Dans la pratique, le MDS est souvent utilisé en psychométrie (analyse de perceptions) et en marketing (distances entre produits obtenus à partir de classements par des consommateurs), mais on trouve des applications dans de très nombreux domaines.

Si la matrice de départ est une matrice de similarité (une similarité est d'autant plus élevée que deux objets sont proches), elle sera automatiquement convertie en matrice de dissimilarité pour la suite des calculs. La conversion s'effectue en soustrayant à la valeur de la diagonale les données de la matrice.

On distingue deux types de MDS en fonction de la nature de la dissimilarité observée :

- **MDS métrique** : les dissimilarités sont considérées comme continues et donnant une information exacte à reproduire le plus fidèlement possible. Différents sous-modèles sont proposés :

- absolu (*absolute MDS*) : les distances obtenues dans l'espace de représentation doivent correspondre le plus fidèlement possible aux distances observées dans la matrice de dissimilarité initiale.

- rapport (*ratio MDS*) : les distances obtenues dans l'espace de représentation doivent correspondre le plus fidèlement possible aux distances observées dans la matrice initiale, à un facteur de proportionnalité près (le facteur étant identique pour tous les couples de distances).

- intervalle (*interval MDS*) : les distances obtenues dans l'espace de représentation doivent correspondre le plus fidèlement possible aux distances observées dans la matrice initiale, à une relation linéaire près (la relation linéaire facteur étant identique pour tous les couples de distances).

- polynomial (*polynomial MDS*) : les distances obtenues dans l'espace de représentation doivent correspondre le plus fidèlement possible aux distances observées dans la matrice initiale, à une relation polynomiale de degré deux près (la relation linéaire facteur étant identique pour tous les couples de distances).

Remarque : le modèle absolu permet de comparer les distances dans l'espace de représentation à celles de l'espace de départ. Les autres modèles présentent l'avantage d'accélérer les calculs.

- **MDS non métrique** : avec ce type de MDS, seul compte l'ordre entre les dissimilarités. Autrement dit l'algorithme MDS ne doit pas essayer de reproduire les dissimilarités, mais seulement la relation d'ordre entre ces dernières. Deux modèles sont possibles :

- ordinal (1) : la relation d'ordre entre les distances dans l'espace de représentation doit correspondre à celle des dissimilarités correspondantes. En cas de dissimilarités de même rang, aucune restriction n'est imposée sur les distances correspondantes. Autrement dit, des dissimilarités de même rang ne doivent pas nécessairement donner des distances égales dans l'espace de représentation.

- ordinal (2) : la relation d'ordre entre les distances dans l'espace de représentation doit correspondre à celle des dissimilarités correspondantes. En cas de dissimilarités de même rang, les distances correspondantes doivent être égales.

Les algorithmes de MDS visent à réduire l'écart entre la matrice des disparités issues des modèles et la matrice des distances obtenues dans la configuration de représentation. Dans le cas du modèle absolu, la disparité est égale à la dissimilarité de la matrice de départ. L'écart est mesuré au travers du Stress dont plusieurs variantes ont été proposées :

- Stress brut :

$$\sigma_r = \sum_{i<j} w_{ij}(D_{ij} - d_{ij})^2,$$

où D_{ij} représente la disparité entre l'individu i et l'individu j, d_{ij} la distance euclidienne entre ces mêmes individus pour la représentation obten, et w_{ij} représente le poids affecté à la proximité ij (par défaut sa valeur est 1).

- Stress standardisé :

$$\sigma_r = \frac{\sum_{i<j} w_{ij}(D_{ij} - d_{ij})^2}{\sum_{i<j} w_{ij} D_{ij}^2}.$$

- Stress 1 de Kruskal :

$$\sigma_r = \sqrt{\frac{\sum_{i<j} w_{ij}(D_{ij} - d_{ij})^2}{\sum_{i<j} w_{ij} d_{ij}^2}}.$$

- Stress 2 de Kruskal :

$$\sigma_r = \sqrt{\frac{\sum_{i<j} w_{ij}(D_{ij} - d_{ij})^2}{\sum_{i<j} w_{ij}(d_{ij} - \bar{d})^2}}.$$

où $\bar{d}$ représente la moyenne des distances sur la représentation.

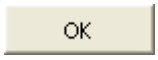
Remarque : pour un nombre de dimensions donné, plus faible est le stress, meilleure est la qualité de la représentation. Par ailleurs, plus le nombre de dimensions est élevé, plus le stress est faible.

Pour savoir si le résultat obtenu est satisfaisant et pour déterminer quel est le nombre de dimensions correct pour représenter fidèlement les données, on peut observer l'évolution du stress avec le nombre de dimensions et identifier à partir de quand le stress se stabilise. Le diagramme de Shepard permet quant à lui d'observer d'éventuelles ruptures dans l'ordination des distances. Plus le graphique est linéaire, meilleure est la représentation. Dans le cas du modèle absolu, pour une représentation idéale, les points doivent être alignés sur la première bissectrice.

Il existe plusieurs algorithmes de MDS dont notamment ALSCAL (Takane *et al*. 1977) et SMACOF (*Scaling by MAjorizing a COnvex Function*) qui minimise le « Stress standardisé » (de Leeuw, 1977). XLSTAT utilise l'algorithme SMACOF.

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT vous permet de sélectionner les données par colonnes ou par plage. Si la flèche est vers la droite, XLSTAT vous permet de sélectionner les données par lignes ou par plage.

Onglet **Général** :

Données : sélectionnez une matrice de similarité ou dissimilarité. Si seule la partie triangulaire inférieure ou supérieure est disponible, le tableau est accepté. Si des différences sont détectées entre les parties inférieure et supérieure de la matrice sélectionnée, XLSTAT vous en avertit, et vous propose de modifier les données (calcul de la moyenne des deux parties) pour pouvoir poursuivre les calculs.

Dissimilarités / Similarités : choisissez l'option correspondant à la nature des données de matrice sélectionnée.

Modèle : choisissez le modèle à utiliser. Voir la partie description pour plus de détails.

Dimensions : entrez les nombres minimum et maximum de dimensions pour l'espace de représentation des objets. L'algorithme sera répété pour toutes les dimensions se trouvant entre ces deux bornes.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés inclus : activez cette option si vous avez inclus les libellés des lignes et des colonnes dans la sélection.

Poids : activez cette option si vous voulez pondérer les données. Vous devez alors sélectionner une matrice de poids (sans sélectionner de libellés pour les lignes ou les colonnes). Si vous n'activez pas cette option, les poids seront tous considérés comme valant 1. Les poids doivent être impérativement supérieurs ou égaux à 0.

Onglet **Options** :

Stress : choisissez le type de stress qui sera utilisé pour la restitution des résultats, sachant que l'algorithme SMACOF minimise le stress brut. Voir la section description pour plus de détails.

Configuration initiale :

- **Aléatoire** : activez cette option pour que XLSTAT génère de manière aléatoire la configuration de départ. Entrez alors le nombre de fois où l'algorithme devra être répété à partir d'une nouvelle configuration initiale générée aléatoirement. Valeur par défaut du nombre de répétitions : 100. Remarque : la configuration affichée dans les résultats correspond à la répétition pour laquelle le meilleur résultat a été trouvé.
- **Définie par l'utilisateur** : activez cette option pour pouvoir ensuite sélectionner une configuration initiale sur la base de laquelle l'algorithme réalisera ensuite l'optimisation.

Conditions d'arrêt :

- **Itérations** : entrez le nombre maximal d'itérations pour l'algorithme SMACOF. L'optimisation du Stress est arrêtée dès que le nombre maximal d'itérations est dépassé. Valeur par défaut : 100.
- **Convergence** : entrez la valeur minimale d'évolution du stress d'une itération à l'autre, qui une fois atteinte permet de considérer que l'algorithme a convergé. Valeur par défaut : 0,00001.

Onglet **Données manquantes** :

Ne pas accepter les valeurs manquantes : activez cette option pour que XLSTAT empêche la poursuite des calculs si des valeurs manquantes sont détectées.

Ignorer les données manquantes : si vous activez cette option, XLSTAT ne tiendra pas compte des proximités correspondant à des données manquantes pour la minimisation du stress.

Onglet **Sorties** :

Distances : activez cette option pour afficher la matrice des distances euclidiennes correspondant à la configuration optimale.

Disparités : activez cette option pour afficher la matrice des disparités correspondant à la configuration optimale.

Distances résiduelles : activez cette option pour afficher la matrice des distances résiduelles correspondant à la différence entre la matrice des distances et la matrice des disparités.

Onglet **Graphiques** :

Evolution du stress : activez cette option pour afficher le graphique d'évolution du stress en fonction du nombre de dimensions de la configuration.

Configuration : activez cette option pour afficher le graphique de représentation de la configuration. Ce graphique n'est affiché que pour la configuration dans un espace de dimension 2 si ce dernier est calculé.

- **Etiquettes** : activez cette option pour afficher les étiquettes des objets.
- **Etiquettes colorées** : activez cette option pour que les étiquettes soient de la même couleur que les points.
- **Diagramme de Shepard** : activez cette option pour afficher le diagramme de Shepard.

Résultats

Stress après minimisation : ce tableau permet de visualiser pour les dimensions étudiées le stress final obtenu, le nombre d'itérations nécessaire et le niveau de convergence atteint. Dans le cas où plusieurs dimensions sont étudiées, un graphique d'évolution du stress en fonction du nombre de dimensions est affiché.

Les résultats qui suivent sont affichés pour chacune des dimensions étudiées.

Configuration : dans ce tableau sont affichées les coordonnées des objets dans l'espace de représentation. Si l'espace est à deux dimensions, une représentation graphique de la configuration est fournie. Si vous disposez de l'outil XLSTAT-3DPlot, vous pouvez aussi visualiser une configuration en trois dimensions.

Distances mesurées dans l'espace de représentation : ce tableau correspond aux distances entre les objets dans l'espace de représentation.

Disparités calculées d'après le modèle : ce tableau fournit les disparités calculées à partir du modèle choisi (absolu, intervalle, …).

Distances résiduelles : ces distances sont la différence entre les dissimilarités de la matrice initiale, et les distances mesurées dans l'espace de représentation.

Tableau de comparaison : ce tableau permet de comparer les dissimilarités, les disparités et les distances, ainsi que les rangs de ces trois mesures pour l'ensemble des combinaisons deux à deux d'objets.

Diagramme de Shepard : ce graphique permet de comparer les disparités et les distances aux dissimilarités. Dans le cas d'un modèle métrique, la représentation est d'autant meilleure que les points sont confondus avec la première bissectrice du plan. Dans le cas d'un modèle non métrique, le modèle est d'autant meilleur que la ligne des dissimilarités/disparités croît régulièrement. Par ailleurs la performance du modèle peut être évaluée en observant si les points (dissimilarité/distance) sont proches des points (dissimilarité/disparité).

Exemple

Un exemple d'utilisation de Multidimensional Scaling est disponible sur le Centre d'aide XLSTAT à l'adresse suivante :

http://www.xlstat.com/demo-mdsf.htm

Bibliographie

Borg I. and Groenen P. (1997). Modern Multidimensional Scaling. Theory and applications. Springer Verlag, New York.

Cox T.C. and Cox M.A.A. (2001). Multidimensional Scaling (2nd edition). Chapman and Hall, New York.

De Leeuw J. (1977). Applications of Convex Analysis to Multidimensional Scaling, in J.R. Barra a.o. (eds.), Recent Developments in Statistics. North Holland Publishing Company, Amsterdam. 133-146.

Heiser W.J. (1991). A general majorization method for least squares multidimensional scaling of pseudodistances that may be negative. *Psychometrika*, **56** (1), 7-27.

Kruskal J.B., Wish M. (1978). Multidimensional Scaling. Sage Publications, London.

Takane Y., Young F. W. and DeLeeuw J. (1977). Nonmetric individual differences multidimensional scaling: an alternating least squares method with optimal scaling features. *Psychometrika*, **42**, 7-67.

Classification k-means

Utilisez la classification k-means pour constituer des groupes homogènes d'objets (classes) sur la base de leur description par un ensemble de variables quantitatives.

Dans cette section :

Description

La classification k-means a été introduite par MacQueen en 1967. D'autres algorithmes similaires ont été développés par Forgey (1965) (centres mobiles) et Friedman (1967).

La classification k-means présente notamment les avantages suivants : un objet peut être affecté à une classe au cours d'une itération puis changer de classe à l'itération suivante, ce qui n'est pas possible avec la classification ascendante hiérarchique pour laquelle une affectation est irréversible.

En multipliant les points de départ et les répétitions on peut explorer plusieurs solutions possibles.

L'inconvénient de cette méthode est qu'elle ne permet pas de découvrir quel peut être un nombre cohérent de classes, ni de visualiser la proximité entre les classes ou les objets.

Les méthodes k-means et CAH sont donc complémentaires.

Remarque : dans le cas où vous souhaiteriez prendre en compte des variables qualitatives pour la classification, il est nécessaire d'effectuer au préalable une analyse des correspondances multiples (ACM) et de considérer les coordonnées des individus sur les axes factoriels obtenus comme de nouvelles variables.

Principe de la méthode k-means

La classification k-means est une méthode itérative qui, quel que soit son point de départ, converge vers une solution. La solution obtenue n'est pas nécessairement la même quel que soit le point de départ. Pour cette raison, on répète en général plusieurs fois les calculs pour ne retenir que la solution la plus optimale pour le critère choisi.

Pour la première itération on choisit un point de départ qui consiste à associer le centre des k classes à k objets (pris au hasard ou non). On calcule ensuite la distance entre les objets et les k centres et on affecte les objets aux centres dont ils sont les plus proches. Puis on redéfinit les centres à partir des objets qui ont été affectés aux différentes classes. Puis on affecte à nouveau les objets en fonction de leur distance aux nouveaux centres. Et ainsi de suite jusqu'à ce que la convergence soit atteinte.

Critères de classification

Plusieurs critères de classification peuvent être utilisés pour parvenir à une solution. XLSTAT propose quatre critères à minimiser.

Trace(W) : la trace de W, matrice d'inertie intra-classe commune (*pooled SSCP matrix*) est le critère le plus classique. Minimiser la trace de W pour un nombre de classes donné, revient à minimiser la variance intra-classe totale, autrement dit à minimiser l'hétérogénéité des groupes. Ce critère est sensible aux effets d'échelle. Si on ne veut pas donner plus de poids à certaines variables plutôt qu'à d'autres, on doit préalablement normaliser les données. Par ailleurs, ce critère tend à produire des classes de même taille.

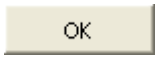
Déterminant(W) : le déterminant de W, matrice de covariance intra-classe commune (*pooled within covariance matrix*) est un critère nettement moins sensible aux effets d'échelle que le critère trace(W). Par ailleurs, la taille des groupes peut être moins homogène qu'avec le critère de la trace.

Wilks lambda : les résultats donnés par la minimisation de ce critère sont identiques à ceux donnés par le déterminant de W. Le critère du lambda de Wilks correspond à la division du déterminant(W) par le déterminant(T) où T est la matrice d'inertie totale. La division par le déterminant de T permet d'avoir un critère toujours compris entre 0 et 1.

Trace(W) / Médiane : si l'on choisit ce critère, le barycentre d'une classe n'est pas le point moyen de la classe, mais le point médian qui correspond à un objet de la classe. L'utilisation de ce critère entraîne des calculs plus longs.

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.

 : cliquez sur ce bouton pour effacer les sélections de données.

 : cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.

 : cliquez sur ce bouton pour changer la façon dont vous chargez les données dans XLSTAT. Si le bouton possède une icône de souris, XLSTAT vous permet de faire une sélection à la souris des données. Si le bouton possède une icône de liste, XLSTAT vous permet de sélectionner les données à partir d'une liste. Si le bouton possède une icône orange, des boutons complémentaires avec un point d'interrogation  sont affichés pour vous permettre d'importer les données à partir de fichiers.

Onglet **Général** :

Tableau observations/variables : sélectionnez un tableau comprenant N objets décrits par P descripteurs. Si des en-têtes de colonnes ont été sélectionnés, veuillez vérifier que l'option « Libellés des variables » est activée.

Indice de dissimilarité : choisissez parmi les trois distances proposées par XLSTAT. La distance euclidienne est la plus utilisée pour la classification k-means et s'applique à la majorité des cas. La distance cosinus est adaptée aux données textuelles. La distance de Jaccard est adaptée aux données qui requièrent une finesse d'analyse importante.

Critère de classification : choisissez le critère de classification (voir la section description pour plus de détails).

Nombre de classes : entrez le nombre de classes qui doivent être créées par l'algorithme. Vous pouvez choisir de faire varier le nombre de classes entre deux valeurs.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des colonnes : activez cette option si la première ligne des données sélectionnées (Tableau observations/variables, libellés des lignes, poids des lignes, poids des colonnes) contient un libellé.

Libellés des lignes : activez cette option si vous voulez utiliser des libellés d'observations pour l'affichage des résultats. Si l'option « Libellés des colonnes » est activée, la première cellule de la sélection doit comprendre un en-tête. Si vous n'activez pas cette option, des libellés seront automatiquement créés (Obs1, Obs2, …).

Poids des colonnes : activez cette option si vous voulez pondérer les colonnes. Si vous n'activez pas cette option, les poids seront tous considérés comme valant 1. Les poids doivent

être impérativement supérieurs ou égaux à 0. Si un en-tête de colonne a été sélectionné, veuillez vérifier que l'option « Libellés des colonnes » est activée.

Poids des lignes : activez cette option si vous voulez pondérer les lignes. Si vous n'activez pas cette option, les poids seront tous considérés comme valant 1. Les poids doivent être impérativement supérieurs ou égaux à 0. Si un en-tête de colonne a été sélectionné, veuillez vérifier que l'option « Libellés des colonnes » est activée.

Onglet **Options** :

Regrouper les lignes : activez cette option si vous voulez créer des classes d'objets correspondant aux lignes et décrits par les données correspondant aux colonnes.

Regrouper les colonnes : activez cette option si vous voulez créer des classes d'objets correspondant aux colonnes et décrits par les données correspondant aux lignes.

Centrer : activez cette option si vous voulez centrer les données avant de commencer les calculs.

Réduire : activez cette option si vous voulez réduire les données avant de commencer les calculs.

Vous pouvez ensuite choisir si la transformation doit être appliquée aux lignes ou aux colonnes.

Résultats dans l'espace d'origine : activez cette option pour afficher les résultats dans l'espace d'origine. Si les options centrer/réduire sont activées, et que cette option n'est pas activée, les résultats sont fournis dans l'espace centré/réduit.

Conditions d'arrêt :

- **Itérations** : entrez le nombre maximal d'itérations pour l'algorithme k-means. Les calculs sont interrompus dès que le nombre maximal d'itérations est dépassé. Valeur par défaut : 500.
- **Convergence** : entrez la valeur minimale d'évolution du critère choisi d'une itération à l'autre, qui une fois atteinte permet de considérer que l'algorithme a convergé. Valeur par défaut : 0,00001.

Partition de départ : utilisez ces options pour choisir la manière dont est déterminée la partition initiale, autrement dit, la façon dont sont affectés les objets aux classes pour la première itération de l'algorithme de classification.

- N **classes d'après l'ordre** : les objets sont affectés aux classes en fonction de leur ordre.
- **Aléatoire** : les objets sont affectés aux classes de manière aléatoire.
- **Définie par l'utilisateur** : les objets sont affectés aux classes suivant une variable indicatrice définie par l'utilisateur. L'utilisateur doit dans ce cas sélectionner une variable indicatrice en colonne contenant autant de lignes que d'objets (avec éventuellement un en-tête), et les classes doivent être définies par des valeurs de 1 à k, où k est le nombre de classes. Si le nombre de classes k varie, l'utilisateur doit entrer autant de colonnes

que de k différents. Si l'option « Libellés des colonnes » est activée, la première cellule de la sélection doit comprendre un en-tête.

- **Définie par les centres** : l'utilisateur doit sélectionner les k centres correspondant aux k classes. Dans le cas où le regroupement est fait sur les lignes, le nombre de lignes définit le nombre de classes à créer et le nombre de colonnes doit être le même que le nombre de colonnes du tableau des données. A contrario, dans le cas où le regroupement est fait sur les colonnes, le nombre de colonnes définit le nombre de classes à créer et le nombre de lignes doit être le même que le nombre de lignes du tableau des données. Si l'option « Libellés des colonnes » est activée, la première cellule de la sélection doit comprendre un en-tête.

- **K++** : cette option définit les centres initiaux en fonction de l'algorithme k-means++ développé par Rafail Ostrovsky, Yuval Rabani, Leonard Schulman et Chaitanya Swamy en 2006. Le premier centre est choisi aléatoirement parmi les observations. Le suivant est choisi parmi les observations en fonction de la distance entre l'observation et le centre. Plus la distance entre l'observation et le centre est grande et plus l'observation a de chance d'être sélectionnée. Les $k-2$ centres restants sont choisis en suivant la même méthode. Cette méthode permet de démarrer à partir de centres sélectionnés de manière homogène dans le jeu de données, ce qui conduit généralement à un nombre d'itérations moins grand et une qualité de partitionnement meilleure. En revanche, sur des données de grande taille et complexes (contenant beaucoup de centres), cet algorithme peut prendre un temps d'exécution non négligeable et il est préférable d'utiliser l'algorithme K||.

- **K||** : cette option définit les centres en fonction de l'algorithme K|| ou "Scalable k-means" développé par Bahman Bahmani, Benjamin Moseley, Andrea Vattani, Ravi Kumar et Sergei Vassilvitskii en 2012. C'est une version modifiée de K++ qui permet d'effectuer le choix des centres initiaux en parallèle. Comme avec K++, le premier centre est choisi aléatoirement parmi les observations, puis à la prochaine itération, $\tilde{k}/2$ observations sont choisies de manière indépendante en fonction de leur distance par rapport au centre. Après un nombre d'itérations déterminé en fonction de la taille du jeu de donnée, les X centres ainsi obtenus sont agrégés en k centres via l'algorithme K++. Cet algorithme possède l'avantage d'être très rapide pour deux raisons : La première étape consiste à ne prendre qu'une partie des observations du jeu de données ce qui facilite grandement le travail de K++ et le choix de manière indépendante des centres à chaque itération permet de paralléliser une grande partie de l'algorithme.

Onglet **Prédiction** :

Prédiction : disponible uniquement si vous effectuez un regroupement sur les lignes. Activez cette option si vous souhaitez sélectionner des données à utiliser en mode prédiction. Si vous activez cette option, vous devez veiller à ce que les données de prédiction soient organisées comme les données d'apprentissage : mêmes variables, même ordre dans les sélections.

Tableau observations/variables : sélectionnez un tableau comprenant les nouveaux objets à prédire et décrits par les mêmes P descripteurs que le tableau d'apprentissage. Si des en-têtes de colonnes ont été sélectionnés, veuillez vérifier que l'option « Libellés des variables » est activée.

Libellés des observations : activez cette option si vous voulez utiliser des libellés d'observations disponibles sur une feuille Excel pour l'affichage des résultats. Si vous n'activez pas cette option, des libellés seront automatiquement créés (PredObs1, PredObs2, …).

Libellés des variables : activez cette option si la première ligne des données sélectionnées (variables, libellés des observations) contient un libellé.

Onglet **Données manquantes** :

Ne pas accepter les valeurs manquantes : activez cette option pour que XLSTAT empêche la poursuite des calculs si des valeurs manquantes sont détectées.

Supprimer les observations : activez cette option pour supprimer les observations comportant des données manquantes.

Estimer les données manquantes : activez cette option pour estimer les données manquantes avant le début des calculs.

- **Moyenne ou mode** : activez cette option pour estimer les données manquantes en utilisant la moyenne (variables quantitatives) ou le mode (variables qualitatives) pour les variables correspondantes.
- **Plus proche voisin** : activez cette option pour estimer les données manquantes d'une observation en recherchant le plus proche voisin de l'observation.

Onglet **Sorties** :

Statistiques descriptives : activez cette option pour afficher les statistiques descriptives pour les variables sélectionnées et pour chaque jeu de données présent (apprentissage, prédiction).

Matrice de corrélation : activez cette option pour afficher un aperçu des corrélations entre les différentes variables sélectionnées pour l'échantillon d'apprentissage.

Bilan de l'optimisation : activez cette option pour la synthèse de l'optimisation.

Barycentres : activez cette option pour afficher les coordonnées des barycentres des classes.

Objets centraux : activez cette option pour afficher les coordonnées de l'objet le plus proche du barycentre de chaque classe.

Contribution (Analyse de la variance) : activez cette option pour afficher un tableau donnant la contribution de chaque variable.

Résultats par classe : activez cette option pour afficher un tableau donnant les statistiques et les objets correspondant à chacune des classes.

Résultats par objet : activez cette option pour afficher un tableau donnant pour chaque objet (dans l'ordre initial des objets) sa classe d'affectation et la distance le séparant du barycentre de sa classe.

- **Corrélations avec les barycentres** : activez cette option pour afficher la corrélation de Pearson entre une observation et le barycentre de sa classe.
- **Coefficient de silhouette** : activez cette option pour afficher le coefficient de silhouette de chaque observation.
    - **Moyenne par classe** : activez cette option pour afficher un tableau donnant le coefficient de silhouette moyen de chaque classe.

Onglet **Graphiques** :

Évolution du critère : activez cette option pour afficher le graphique d'évolution du critère choisi.

Profil des classes : activez cette option pour afficher un graphique permettant de comparer les moyennes des différentes classes créées.

Coefficient de silhouette : activez cette option pour afficher le graphique associé aux coefficients de silhouette de chaque observation de l'échantillon.

Coefficient de silhouette (Moyennes) : activez cette option pour afficher le coefficient de silhouette moyen de chaque classe.

Résultats

Statistiques descriptives : le tableau des statistiques descriptives présente des statistiques simples pour toutes les variables sélectionnées. Le nombre de valeurs manquantes, le nombre de valeurs non manquantes, la moyenne, l'écart-type sont affichés pour les variables quantitatives. Pour les variables qualitatives, les catégories avec leur fréquence respective et pourcentage sont affichées.

Matrice de corrélation : ce tableau est affiché afin de vous permettre d'avoir un aperçu des corrélations entre les différentes variables sélectionnées.

Évolution de l'inertie intra-classe : si vous avez choisi un nombre de classes entre deux bornes distinctes, XLSTAT affiche dans un premier temps l'évolution de l'inertie intra-classe, qui diminue mathématiquement lorsque le nombre de classes augmente. Si les données sont distribuées de manière homogène, la décroissance est linéaire. S'il y a réellement une structure de groupes, un coude sera observé pour le nombre de classes pertinent.

Évolution du coefficient de silhouette : si vous avez choisi un nombre de classes entre deux bornes distinctes et que l'option *Coefficient de Silhouette* est activée, un tableau avec le graphique associé montre l'évolution du coefficient de silhouette pour chaque k. Il est possible de choisir le nombre de classes pertinent en prenant le nombre de classes avec le coefficient de silhouette le plus proche de 1.

Bilan de l'optimisation : dans ce tableau est affichée l'évolution de la variance intra-classe. Si plusieurs répétitions ont été demandées, les résultats sont affichés pour chaque répétition. De plus, les résultats de la meilleure répétition sont affichés en gras.

Statistiques pour chaque itération : dans ce tableau est affichée l'évolution des diverses statistiques calculées au fur et à mesure des itérations de la répétition ayant donné le résultat optimal pour le critère choisi. Si l'option correspondante est activée dans l'onglet Graphiques, un graphique présentant l'évolution du critère choisi au fur et à mesure des itérations est affiché.

Décomposition de l'inertie pour la classification optimale : dans ce tableau sont affichées l'inertie intra-classe, l'inertie inter-classes et l'inertie totale.

Remarque : si les données sont centrées/réduites (option de l'onglet Options) les résultats pour le bilan de l'optimisation, les statistiques pour chaque itération et le tableau de la décomposition de l'inertie sont calculés dans l'espace centré-réduit. En revanche, les résultats qui suivent sont affichés dans l'espace d'origine si l'option « Résultats dans l'espace d'origine » est activée.

Barycentres initiaux des classes : dans ce tableau sont affichées les coordonnées des barycentres qui ont été calculées grâce à la partition initiale aléatoire ou grâce aux algorithmes K|| et K++. Dans le cas où les centres ont été définis par l'utilisateur, ce sont les centres définis qui sont affichés.

Barycentres des classes : dans ce tableau sont affichées les coordonnées des barycentres des classes pour les différents descripteurs.

Distances entre les barycentres des classes : dans ce tableau sont affichées les distances entre les barycentres des classes pour les différents descripteurs.

Objets centraux : dans ce tableau sont affichées, pour chaque classe, les coordonnées de l'objet le plus proche du barycentre de la classe.

Distances entre les objets centraux : dans ce tableau sont affichées les distances entre les objets centraux des classes pour les différents descripteurs.

Résultats par classe : les statistiques descriptives des classes (nombre d'objets, somme des poids, variance intra-classe, distance minimale au barycentre, distance maximale au barycentre, distance moyenne au barycentre) sont affichées dans la première partie du tableau. Dans la seconde partie sont affichés les objets.

Résultats par objet : dans ce tableau est indiquée, pour chaque objet, sa classe d'affectation dans l'ordre initial des objets. * **Distance au barycentre** : cette colonne indique la distance entre une observation et le barycentre de sa classe. * **Corrélations avec les barycentres** : cette colonne indique la corrélation de Pearson entre une observation et le barycentre de sa classe. * **Observation bruitée** : cette colonne indique quelle observation est un bruit par un "Oui" en gras. * **Coefficients de silhouette** : cette colonne indique le coefficient de silhouette de chaque observation.

Le coefficient de silhouette permet de mesurer le bon classement d'une observation dans sa classe. Il se calcule comme suit : $$ s(i) = \frac{b(i) - a(i)}{max(a(i),b(i))}, $$ avec $a(i)$ la distance moyenne du point i par rapport aux autres points de sa classe et $b(i)$ la distance moyenne du point i par rapport aux points de la classe la plus proche. Le coefficient de silhouette varie entre -1 et 1 et plus sa valeur est proche de 1 plus l'observation est considérée comme bien classée.

Coefficients de silhouette (Moyenne par classe) : ce tableau et le graphique associé affichent le coefficient de silhouette moyen de chaque classe et en dernière ligne le coefficient de silhouette de la classification optimale (moyenne des moyennes par classe).

Contribution (Analyse de la variance) : ce tableau permet de connaître les variables qui contribuent le plus à la séparation des classes en effectuant une ANOVA.

Profil des classes : ce graphique permet de comparer visuellement les moyennes des différentes classes créées.

Exemple

Un exemple de classification k-means est disponible sur le Centre d'aide XLSTAT à l'adresse suivante :

http://www.xlstat.com/demo-cluster2f.htm

Bibliographie

Arabie P., Hubert L.J. and De Soete G. (1996). Clustering and Classification. Wold Scientific, Singapore.

Everitt B.S., Landau S. and Leese M. (2001). Cluster Analysis (4th edition). Arnold, London.

Forgey E. (1965). Cluster analysis of multivariate data: efficiency versus interpretability of classification. *Biometrics*, **21**, 768.

Friedman H.P. and Rubin J. (1967). On some invariant criteria for grouping data. Journal of the American Statistical Association, 62, 1159-1178.

Jobson J.D. (1992). Applied Multivariate Data Analysis. Volume II: Categorical and Multivariate Methods. Springer-Verlag, New York, 483-568.

MacQueen J. (1967). Some method for classification and analysis of multivariate observations. In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 281-297.

Saporta G. (1990). Probabilités, Analyse des Données et Statistique. Technip, Paris, 251-260.

Classification Ascendante Hiérarchique (CAH)

Utilisez la classification ascendante hiérarchique pour constituer des groupes homogènes d'objets (classes) sur la base de leur description par un ensemble de variables, ou à partir d'une matrice décrivant la similarité ou la dissimilarité entre les objets.

Dans cette section :

Description

Principe de la CAH

La classification ascendante hiérarchique (CAH) est une méthode de classification itérative dont le principe est simple.

Nous commençons par calculer la dissimilarité entre les N objets. Puis les deux objets dont le regroupement minimise un critère d'agrégation donné, sont regroupés créant ainsi une classe comprenant ces deux objets. Nous calculons ensuite la dissimilarité entre cette classe et les $N-2$ autres objets en utilisant le critère d'agrégation. Puis les deux objets ou classes d'objets dont le regroupement minimise le critère d'agrégation sont à leur tour regroupés. Nous continuons ainsi jusqu'à ce que tous les objets soient regroupés.

Ces regroupements successifs produisent un arbre binaire de classification (dendrogramme), dont le haut correspond à la classe regroupant l'ensemble des individus. Ce dendrogramme représente une hiérarchie de partitions.

On peut alors choisir une partition en tronquant l'arbre à un niveau donné, le niveau dépendant soit des contraintes de l'utilisateur (l'utilisateur sait combien de classes il veut obtenir), soit de critères plus objectifs.

Similarités et dissimilarités

La mesure de la proximité entre deux objets peut se faire en mesurant à quel point ils sont semblables (similarité) ou dissemblables (dissimilarité). Si l'utilisateur choisit une similarité, XLSTAT la convertit ensuite en dissimilarité, car l'algorithme de la CAH utilise les dissimilarités. La conversion pour chaque couple d'objets consiste à prendre la similarité maximale pour l'ensemble des couples, et de lui soustraire ensuite la similarité du couple en question.

Les indices de **similarité** proposés pour des calculs à partir de données quantitatives sont les suivants : Cooccurrence, Cosinus, Covariance (n-1), Covariance (n), Indice de Dice, Inertie, Indice de Jaccard, Coefficient de corrélation de Kendall, Indice de Kulczinski, Indice d'Ochiai, Coefficient de corrélation de Pearson, Phi de Pearson, Similarité générale, Indice de Rogers & Tanimoto, Indice de Sokal & Michener (*simple matching coefficient*), Indice de Sokal & Sneath(1), Indice de Sokal & Sneath(2), Coefficient de corrélation de Spearman, Corrélations carrées.

Les indices de **dissimilarité** proposés pour des calculs à partir de données quantitatives sont les suivants : Distance de Bhattacharya, Distance de Bray et Curtis, Distance de Canberra, Distance de Chebychev, Distance du Khi², Métrique du Khi², Distance de la corde, Distance de la corde au carré, Indice de Dice, Distance euclidienne, Distance géodésique, Indice de Jaccard, Dissimilarité de Kendall, Indice de Kulczinski, Distance de Mahalanobis, Distance de Manhattan, Indice d'Ochiai, Dissimilarité de Pearson, Phi de Pearson, Dissimilarité générale, Indice de Rogers & Tanimoto, Indice de Sokal & Michener, Indice de Sokal & Sneath(1), Indice de Sokal & Sneath(2), Dissimilarité de Spearman, Corrélation carrées.

Remarque : certains indices sont utilisés sur des données binaires (voir la section sur les matrices de similarité/dissimilarité).

Qualité d'une classification hiérarchique

Afin de déterminer la qualité d'une classification hiérarchique, nous pouvons nous aider du coefficient de corrélation cophénétique basé sur la notion de distance cophénétique.

La distance cophénétique entre deux observations est estimée à partir du dendrogramme issu de la classification. Elle est égale à la hauteur du dendrogramme à laquelle les deux observations se retrouvent rassemblées dans la même classe.

Toutes les distances ainsi créées sont rassemblées dans une matrice symétrique de distances cophénétiques.

La corrélation cophénétique est égale au coefficient de corrélation de Pearson entre la matrice de dissimilarité ayant servie pour la classification et la matrice de distances cophénétiques. Plus la corrélation cophénétique est proche de 1, meilleure est la classification.

Méthodes d'agrégation

Pour calculer la dissimilarité entre deux groupes d'objets A et B, différentes stratégies sont possibles. XLSTAT propose les méthodes suivantes :

Lien simple : la dissimilarité entre A et B est la dissimilarité entre l'objet de A et l'objet de B les plus ressemblants. L'agrégation par le lien simple a tendance à contracter l'espace des données et à écraser les niveaux des paliers du dendrogramme. Comme la dissimilarité entre deux éléments de A et de B suffit à relier A et B, ce critère peut conduire à relier des classes très allongées (effet de chaînage) alors qu'elles ne sont pas homogènes.

Lien complet : la dissimilarité entre A et B est la plus grande dissimilarité entre un objet de A et un objet de B. L'agrégation par le lien complet a tendance à dilater l'espace des données et produit des classes compactes.

Lien moyen : la dissimilarité entre A et B est la moyenne des dissimilarités entre les objets de A et les objets de B. L'agrégation selon le lien moyen est un bon compromis entre les critères précédents et respecte assez bien les propriétés de l'espace des données.

Lien proportionnel : la dissimilarité moyenne entre les objets de A et de B est calculée comme une somme de dissimilarités pondérée de telle sorte qu'un poids égal soit attribué aux deux groupes. Comme le lien moyen, ce critère respecte assez bien les propriétés de l'espace des données.

Lien flexible : ce critère fait intervenir un paramètre bêta variant dans l'intervalle $[-1, +1[$ qui permet de générer une famille de critères d'agrégation. Pour bêta = 0, il s'agit du lien proportionnel. Quand bêta est proche de 1, nous obtenons un fort effet de chaînage, mais à mesure que bêta décroît et devient négatif, nous obtenons une dilatation de plus en plus forte.

Méthode de Ward : il s'agit de la méthode la plus utilisée. Nous agrégeons deux groupes de sorte que l'augmentation de l'inertie intra-classe soit la plus petite possible, afin que les classes restent homogènes. Ce critère, proposé notamment par Ward (1963), ne peut s'utiliser que dans le cas des distances quadratiques, c'est-à-dire ici, dans le cas de la distance euclidienne et de la distance du khi².

Tronquer le dendrogramme

Couper ou tronquer le dendrogramme permet d'obtenir une partition après avoir exécuté une Classification Ascendante Hiérarchique. XLSTAT propose 7 méthodes pour couper le dendrogramme, dont 5 méthodes qui coupent le dendrogramme en une partition suivant un critère spécifique.

Trois indices largement utilisés sont proposés : l'indice de Hartigan, le coefficient de Silhouette et l'indice de Calinski et Harabasz. Tout d'abord, XLSTAT vous demande de saisir un intervalle de nombre de classes k au sein duquel les indices préconiseront le nombre de classes à choisir. Ensuite, pour chaque nombre de classes compris dans cet intervalle, une coupure se fait dans le dendrogramme permettant la création d'une partition en k classes. Enfin, les indices sont calculés grâce à la partition obtenue :

- *Indice de Hartigan* : $$ H(k) = (\frac{W_k}{W_{k+1}}-1)(n-k-1), $$ où n est le nombre d'objets, k le nombre de classes, W_k et W_{k+1} sont la somme des carrés intra-classes pour une partition en k classes et de $k + 1$ classes.

- *Coefficient de Silhouette* : $$ S(k) = \frac{1}{n}\sum_ { i = 1} ^ { n }\frac{b(i) - a(i)} {max(a(i),b(i))}, $$

avec $a(i)$ la distance moyenne du point i par rapport aux autres points de sa classe et $b(i)$ la distance moyenne du point i par rapport aux points de la classe la plus proche.

- *Indice de Calinski et Harabasz* : $$ CH(k) = \frac{B_k(n-k)}{W_k(k-1)} , $$ où B_k est la somme des carrés inter-classes pour une partition en k classes.

Ces indices aident à choisir automatiquement le meilleur nombre de classes k. Si vous utilisez le coefficient de Silhouette ou l'indice de Calinski et Harabasz pour tronquer le dendrogramme,

alors c'est le k du plus grand indice qui est choisi. Si vous utilisez l'indice de Hartigan, alors c'est le k donnant la plus grande valeur $H(k-1) - H(k)$ qui est choisi.

Remarque : dans le cas où vous n'avez pas choisi la combinaison distance Euclidienne/critère de Ward, une adaptation des indices d'Hartigan et Calinski Harabasz élaborée par nos équipes est proposée.

Deux autres méthodes permettent de tronquer le dendrogramme par rapport à l'évolution de l'entropie ou de l'inertie entre chaque niveau. Le dendrogramme sera coupé entre les deux niveaux qui donnent la plus grande évolution.

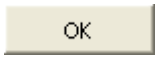
Enfin, vous pouvez soit définir arbitrairement le niveau auquel le dendrogramme doit être tronqué (auquel correspond donc un nombre de classes), soit directement fixer le nombre de classes.

Consolidation avec k-means

Il est possible d'utiliser l'algorithme de k-means après une Classification Ascendante Hiérarchique (CAH), dans le cas où vous avez choisi la distance Euclidienne ou la distance de Jaccard. En effet, k-means peut être exécuté en utilisant la partition générée par la troncature en tant que partition initiale de l'algorithme. Cette méthode permet d'avoir une partition comportant une inertie intra-classe plus faible (ou égale dans le cas où k-means ne trouve pas de meilleure partition). En d'autres termes, l'utilisation consécutive des deux algorithmes, CAH et k-means, permet d'obtenir une partition de meilleure qualité.

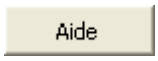
Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.



: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.

 : cliquez sur ce bouton pour changer la façon dont vous chargez les données dans XLSTAT. Si le bouton possède une icône de souris, XLSTAT vous permet de faire une sélection à la souris des données. Si le bouton possède une icône de liste, XLSTAT vous permet de sélectionner les données à partir d'une liste. Si le bouton possède une icône orange, des boutons complémentaires avec un point d'interrogation  sont affichés pour vous permettre d'importer les données à partir de fichiers.

Onglet **Général** :

Le champ principal de saisie des données vous permet de sélectionner alternativement deux types de tableaux :

Tableau observations/variables / Matrice de proximité : choisissez l'option qui correspond au format de vos données, puis sélectionnez les données. Dans le cas de l'option **Tableau observations/variables**, sélectionnez un tableau comprenant N objets décrits par P descripteurs quantitatifs. Dans le cas d'une **matrice de proximité** sélectionnez une matrice carrée donnant les proximités entre les objets. Si des en-têtes de colonnes ont été sélectionnés, veuillez vérifier que l'option « Libellés des colonnes » est activée. Dans le cas d'une matrice de proximité, si les libellés des colonnes sont sélectionnés, ceux des lignes doivent l'être aussi.

Type de proximité : **similarités / dissimilarités** : choisissez le type de proximité à utiliser. Le type de données et le type de proximité déterminent la liste des indices possibles pour le calcul de la matrice de proximité.

Méthode d'agrégation : choisissez la méthode d'agrégation (voir la section description pour plus de détails).

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des colonnes : activez cette option si la première ligne des données sélectionnées (Tableau observations/variables, libellés des lignes, poids des lignes, poids des colonnes) contient un libellé. Dans le cas où la sélection est une matrice de proximité, si cette option est activée, la première colonne doit aussi comprendre le libellé des objets.

Libellés des lignes : activez cette option si vous voulez utiliser des libellés d'observations pour l'affichage des résultats. Si l'option « Libellés des colonnes » est activée, la première cellule de la sélection doit comprendre un en-tête. Si vous n'activez pas cette option, des libellés seront automatiquement créés (Obs1, Obs2, …).

Poids des lignes : activez cette option si vous voulez pondérer les lignes. Si vous n'activez pas cette option, les poids seront tous considérés comme valant 1. Les poids doivent être impérativement supérieurs ou égaux à 0. Si un en-tête de colonne a été sélectionné, veuillez vérifier que l'option « Libellés des colonnes » est activée.

Poids des colonnes : activez cette option si vous voulez pondérer les colonnes. Si vous n'activez pas cette option, les poids seront tous considérés comme valant 1. Les poids doivent être impérativement supérieurs ou égaux à 0. Si un en-tête de colonne a été sélectionné, veuillez vérifier que l'option « Libellés des colonnes » est activée.

Onglet **Options** :

Regrouper les lignes : activez cette option si vous voulez créer des classes d'objets correspondant aux lignes et décrits par les données correspondant aux colonnes.

Regrouper les colonnes : activez cette option si vous voulez créer des classes d'objets correspondant aux colonnes et décrits par les données correspondant aux lignes.

Centrer : activez cette option si vous voulez centrer les données avant de commencer les calculs.

Réduire : activez cette option si vous voulez réduire les données avant de commencer les calculs.

Vous pouvez ensuite choisir si la transformation doit être appliquée aux lignes ou aux colonnes.

Résultats dans l'espace d'origine : activez cette option pour afficher les résultats dans l'espace d'origine. Si les options centrer/réduire sont activées, et que cette option n'est pas activée, les résultats sont fournis dans l'espace centré/réduit.

Variances intra-classe : activez cette option pour sélectionner des variances intra-classe. Cette option est active seulement si les poids des objets ont été sélectionnés (poids des observations dans le cas d'une classification sur les lignes, poids des variables dans le cas d'une classification sur les colonnes). Cette option peut être utilisée si les objets ont déjà été classés par une autre méthode de classification mais que vous voulez utiliser une méthode comme le **Lien moyen** pour classer les groupes obtenus précédemment. Si un en-tête de colonne a été sélectionné, veuillez vérifier que l'option « Libellés des colonnes » est activée.

Troncature : activez cette option si vous voulez que XLSTAT définisse automatiquement une troncature, et donc le nombre de classes à retenir en utilisant une de ces 5 méthodes : **indice de Hartigan**, **indice de Silhouette**, **indice de Calinski et Harabasz**, **Entropie** ou **Inertie**. Vous pouvez aussi définir vous-même le **nombre de classes** à créer (vous pouvez choisir de faire varier le nombre de classes entre deux valeurs), ou le **niveau** auquel le dendrogramme doit être tronqué.

Consolidation : activez cette option si vous voulez exécuter un algorithme de k-means en utilisant la partition générée en tant que partition initiale (voir la section description pour plus de détails). La partition initiale ainsi que la partition obtenue après la consolidation sont affichées dans le tableau des Résultats par objets.

- **Conditions d'arrêt :**

- **Itérations** : entrez le nombre maximal d'itérations pour l'algorithme k-means. Les calculs sont interrompus dès que le nombre maximal d'itérations est dépassé. Valeur par défaut : 500.

- **Convergence** : entrez la valeur minimale d'évolution du critère choisi d'une itération à l'autre, qui une fois atteinte permet de considérer que l'algorithme a convergé. Valeur par défaut : 0,00001.

Onglet **Données manquantes** :

Ne pas accepter les valeurs manquantes : activez cette option pour que XLSTAT empêche la poursuite des calculs si des valeurs manquantes sont détectées.

Supprimer les observations : activez cette option pour supprimer les observations comportant des données manquantes.

Estimer les données manquantes : activez cette option pour estimer les données manquantes avant le début des calculs.

- **Moyenne ou mode** : activez cette option pour estimer les données manquantes en utilisant la moyenne (variables quantitatives) ou le mode (variables qualitatives) pour les variables correspondantes.

- **Plus proche voisin** : activez cette option pour estimer les données manquantes d'une observation en recherchant le plus proche voisin de l'observation.

Onglet **Sorties** :

Statistiques descriptives : activez cette option pour afficher les statistiques descriptives pour les variables sélectionnées.

Matrice de corrélation : activez cette option pour afficher un aperçu des corrélations entre les différentes variables sélectionnées pour l'échantillon d'apprentissage.

Matrice de proximité : activez cette option pour afficher la matrice de proximité.

Statistiques des nœuds : activez cette option pour afficher les statistiques des nœuds du dendrogramme.

Matrice de distances cophénétiques : activez cette option pour afficher la matrice de distances cophénétiques.

Corrélation cophénétique : activez cette option pour afficher le coefficient de corrélation cophénétique.

Décomposition de l'inertie : activez cette option pour afficher le tableau donnant l'inertie intra-classe, l'inertie inter-classes et l'inertie totale.

Résultats par objet : activez cette option pour afficher un tableau donnant pour chaque objet (dans l'ordre initial des objets) sa classe d'affectation et la distance le séparant du barycentre de sa classe.

- **Corrélations avec les barycentres** : activez cette option pour afficher la corrélation de Pearson entre une observation et le barycentre de sa classe.
    - **Observation bruitée** : activez cette option pour afficher une colonne indiquant quelle observation est bruitée. Une observation est bruitée si la corrélation avec le barycentre de sa classe est inférieure au seuil choisi.
- **Coefficient de silhouette** : activez cette option pour afficher le coefficient de silhouette de chaque observation.
    - **Moyenne par classe** : activez cette option pour afficher un tableau donnant le coefficient de silhouette moyen de chaque classe.

Résultats par classe : activez cette option pour afficher un tableau donnant les statistiques et les objets correspondant à chacune des classes.

Barycentres : activez cette option pour afficher les coordonnées des barycentres des classes.

Objets centraux : activez cette option pour afficher les coordonnées de l'objet le plus proche du barycentre de chaque classe.

Onglet **Graphiques** :

Diagramme des niveaux : activez cette option pour afficher le diagramme des niveaux permettant d'observer l'impact des regroupements successifs.

Dendrogramme : activez cette option pour afficher le dendrogramme.

- **Horizontal** : choisissez cette option pour afficher un dendrogramme horizontal.
- **Vertical** : choisissez cette option pour afficher un dendrogramme vertical.
- **Complet** : activez cette option pour afficher le dendrogramme complet (tous les objets sont représentés).
- **Tronqué** : activez cette option pour afficher le dendrogramme tronqué (le dendrogramme commence au niveau de la troncature).
- **Étiquettes** : activez cette option pour afficher les libellés des objets (dendrogramme complet) ou des classes (dendrogramme tronqué) sur le dendrogramme.
- **Couleurs** : activez cette option pour utiliser des couleurs pour représenter les différents groupes sur le dendrogramme complet.
- **Colorer par groupe** : activez cette option pour représenter des groupes prédéfinis sur le dendrogramme en colorant les étiquettes selon le groupe.

Coefficient de silhouette : activez cette option pour afficher le graphique associé aux coefficients de silhouette de chaque observation de l'échantillon.

Coefficient de silhouette (Moyennes) : activez cette option pour afficher le coefficient de silhouette moyen de chaque classe.

Profil des classes : activez cette option pour afficher un graphique permettant de comparer les moyennes des différentes classes créées.

Résultats

Statistiques descriptives : le tableau des statistiques descriptives présente des statistiques simples pour toutes les variables sélectionnées. Le nombre de valeurs manquantes, le nombre de valeurs non manquantes, la moyenne, l'écart-type sont affichés pour les variables quantitatives.

Matrice de corrélation : ce tableau est affiché afin de vous permettre d'avoir un aperçu des corrélations entre les différentes variables sélectionnées.

Matrice de proximité : dans ce tableau sont affichées les proximités entre les objets pour l'indice choisi.

Statistiques des nœuds : dans ce tableau sont affichées les informations concernant les nœuds successifs du dendrogramme. Le premier nœud a pour indice le nombre d'objets augmenté de 1. Ainsi, il est aisé de repérer à quel moment un objet ou un groupe d'objets est regroupé avec un autre objet ou groupe d'objets au niveau d'un nouveau nœud dans le dendrogramme.

Diagramme des niveaux : dans ce tableau sont affichées les statistiques des nœuds du dendrogramme.

Dendrogrammes : le dendrogramme complet permet de visualiser le regroupement progressif des objets. Si une troncature est demandée, un trait en pointillé marque le niveau auquel est effectuée la troncature. Le dendrogramme tronqué permet de visualiser les classes après la troncature.

Matrice des distances cophénétiques : ce tableau indique la distance cophénétique entre chaque observation.

Évolution du coefficient de silhouette : si vous avez choisi un nombre de classes entre deux bornes distinctes et que l'option *Coefficient de Silhouette* est activée, un tableau avec le graphique associé montre l'évolution du coefficient de silhouette pour chaque k. Il est possible de choisir le nombre de classes pertinent en prenant le nombre de classes avec le coefficient de silhouette le plus proche de 1.

Évolution des indices : ce tableau est affiché si vous avez choisi de tronquer le dendrogramme en utilisant l'indice de **Hartigan (H)**, le coefficient de **Silhouette** ou l'indice de **Calinski et Harabasz**. Il indique les valeurs de chaque indice et la valeur $H(k-1) - H(k)$ pour chaque nombre de classes. La valeur en gras correspond à la plus grande valeur qui a permis de déterminer le nombre de classes selon l'indice utilisé.

Note : Une attention particulière doit être apportée dans les tableaux de résultats qui suivent. En effet, la notion de barycentre est liée à la distance euclidienne en général. Les résultats (hormis le calcul du coefficient de Silhouette basé sur la matrice de distance) sont obtenus avec la distance Euclidienne. XLSTAT permet néanmoins d'obtenir ces résultats avec la distance de Jaccard.

Évolution de la variance intra-classe : si vous avez choisi un nombre de classes entre deux bornes distinctes, XLSTAT affiche l'évolution de la variance intra-classe, qui diminue mathématiquement lorsque le nombre de classes augmente. Si les données sont distribuées de manière homogène, la décroissance est linéaire. S'il y a réellement une structure de groupes, un coude sera observé pour le nombre de classes pertinent.

Décomposition de l'inertie : dans ce tableau sont affichées l'inertie intra-classe, l'inertie inter-classes et l'inertie totale.

Barycentres des classes : dans ce tableau sont affichées les coordonnées des barycentres des classes pour les différents descripteurs.

Distances entre les barycentres des classes : dans ce tableau sont affichées les distances euclidiennes entre les barycentres des classes pour les différents descripteurs.

Objets centraux : dans ce tableau sont affichées pour chaque classe les coordonnées de l'objet le plus proche du barycentre de la classe.

Distances entre les objets centraux : dans ce tableau sont affichées les distances euclidiennes entre les objets centraux des classes pour les différents descripteurs.

Résultats par classe : les statistiques descriptives des classes (nombre d'objets, somme des poids, variance intra-classe, distance minimale au barycentre, distance maximale au barycentre, distance moyenne au barycentre) sont affichées dans la première partie du tableau. Les objets sont affichés dans la seconde partie.

Résultats par objet : dans ce tableau est indiquée, pour chaque objet, sa classe d'affectation dans l'ordre initial des objets. * **Classe** : cette colonne donne la partition finale. Il s'agit soit de la partition générée après la coupure du dendrogramme, soit de la partition obtenue avec l'algorithme de k-means si vous avez activé l'option de consolidation. * **Classe (avant consolidation)** : cette colonne donne la partition initiale générée par le dendrogramme tronqué. Cette colonne est affichée dans le cas où vous avez coché l'option de consolidation. * **Distance au barycentre** : cette colonne indique la distance entre une observation et le barycentre de sa classe. * **Corrélations avec les barycentres** : cette colonne indique la corrélation de Pearson entre une observation et le barycentre de sa classe. *NB : il est préférable d'interpréter la corrélation dans un espace centré-réduit. En effet, des conclusions différentes peuvent être amenées entre la corrélation au barycentre et la distance au barycentre si les données ne sont pas à la même échelle. Enfin, quelle que soit la distance choisie, un barycentre est calculé avec la distance euclidienne.* * **Observation bruitée** : cette colonne indique quelle observation est un bruit par un "Oui" en gras. * **Coefficients de silhouette** : cette colonne indique le coefficient de silhouette de chaque observation.

Coefficients de silhouette (Moyenne par classe) : ce tableau et le graphique associé affichent le coefficient de silhouette moyen de chaque classe et en dernière ligne le coefficient de silhouette de la classification optimale (moyenne des moyennes par classe).

Profil des classes : ce graphique permet de comparer visuellement les moyennes des différentes classes créées.

Exemple

Un exemple de Classification Ascendante Hiérarchique est disponible sur le Centre d'aide XLSTAT à l'adresse suivante :

http://www.xlstat.com/demo-clusterf.htm

Bibliographie

Arabie, P., Hubert, L., & De Soete, G. (Eds.). (1996). Clustering and classification. World Scientific.

Arabie, P., Hubert, L., & De Soete, G. (Eds.). (1996). *Clustering and classification*. World Scientific.

Everitt B.S., Landau S. and Leese M. (2001). Cluster analysis (4th edition). Arnold, London.

Jobson J.D. (1992). Applied Multivariate Data Analysis. Volume II: Categorical and Multivariate Methods. Springer-Verlag, New York, 483-568.

Legendre P. and Legendre L. (1998). Numerical Ecology. Second English Edition. Elsevier, Amsterdam.

Saporta G. (1990). Probabilités, Analyse des Données et Statistique. Technip, Paris, 251-260.

Ward J.H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, **58**, 238-244.

Modèles de mélange gaussiens

Cet outil permet d'ajuster des modèles de mélange gaussien à des données continues multidimensionnelles. L'une des applications de ces modèles est la classification des données.

Dans cette section :

Description

Le modèle de mélange gaussien permet la modélisation de données continues par une somme de distributions gaussiennes. Généralement, ces modèles sont utilisés en classification et l'on considère alors que chaque gaussienne représente un groupe.

Le modèle de mélange

On note $x = \{x_1, \ldots, x_n\}$ un vecteur de n variables avec $x_i \in \mathbb{R}^d$. On suppose que chaque X_i est distribuée selon une loi de probabilité de densité f :

$$f(x_i; \theta) = \sum_{k=1}^{K} \pi_k h(x_i; \nu_k)$$

où π_k est la proportion du groupe k ($\forall k \in \{1, \ldots, K\}, 0 < \pi_k < 1 \text{et} \sum_{k=1}^{K} \pi_k = 1$) et θ l'ensemble des paramètres du modèle. La fonction $h(.; \nu_k)$ représente une distribution de probabilité de dimension d et de paramètre ν_k. Par exemple, pour les modèles de mélange gaussiens, h est une distribution gaussienne de moyenne μ_k et de variance Σ_k, et $\nu_k = (\mu_k, \Sigma_k)$.

Pour une distribution de mélange, il est intéressant de remarquer qu'il existe un vecteur de labels $z = \{z_1, \cdots, z_n\}$avec $z_i = \{z_{i1}, \ldots, z_{iK}\}$ défini tel que :

$$\begin{cases} z_{ik} = 1 \text{ si } x_i \text{ provient du k-ieme composant du melange} \\ z_{ik} = 0 \text{ sinon} \end{cases}$$

Ce vecteur de labels est souvent inconnu et dans un contexte de classification ou d'estimation de densité, le but principal est d'estimer chaque z_i.

Inférence des paramètres du modèle

La présence des variables latentes z dans les modèles de mélange ne permet pas d'estimer les paramètres en maximisant directement la log-vraisemblance. Cette maximisation peut être réalisée via des algorithmes itératifs tels que l'algorithme EM (Dempster et al., 1977) ou sa version stochastique proposée dans McLachlan et Peel (2000), appelée SEM.

Une fois les paramètres estimés, le vecteur des labels s'obtient directement en affectant chaque observation x_i au composant dont la probabilité d'appartenance est la plus élevée. Cette probabilité $\hat{\tau}_{ik}$ est définie par :

$$\hat{\tau}_{ik} = \tau_k(x_i; \hat{\theta}) = \frac{\hat{\pi}_k h(x_i; \hat{\nu}_k)}{\sum_{j=1}^{K} \hat{\pi}_j h(x_i; \hat{\nu}_j)}$$

Dans un contexte de classification, Celeux et Govaert (1992) ont proposé l'algorithme CEM (Classification EM) qui est une version classifiante de l'algorithme EM. Contrairement aux algorithmes EM et SEM, l'algorithme CEM ne cherche pas à maximiser la log-vraisemblance mais la quantité $\sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \log[\pi_k h(x_i; \nu_k)]$.

Sélection de modèle (choix du nombre de composants)

Le nombre de composants d'un modèle de mélange est souvent inconnu en pratique. Plusieurs critères tels que le BIC (Bayesian Information Criterion, Schwarz (1978)) ou l'AIC (Akaike Information Criterion, Akaike (1974)) peuvent être utilisés. Ces critères reposent sur une pénalisation de la log-vraisemblance observée $L(x; \theta)$. En 2000, Biernacki *et al*. ont proposé le critère ICL (Integrated Completed Likelihood) qui se base sur une pénalisation de la log-vraisemblance complète $L(x, z; \theta)$. Ce critère peut être écrit comme le BIC pénalisé par un terme d'entropie $-\sum_{i=1}^{n} \sum_{k=1}^{K} \hat{z}_{ik} \log \tau_{ik}$.

Enfin, une estimation du nombre de composants peut être obtenue en mesurant la capacité d'un modèle donné à fournir des groupes bien séparés. Ceci peut être réalisé à l'aide du critère NEC (Normalized Entropy Criterion) proposé par Celeux et Soromengo (1996) :

$$NEC_k = \frac{E_k}{L_k - L_1}$$

où E_k est l'entropie du modèle à k composants et L_k sa log-vraisemblance complète (calculée à l'aide du maximum de vraisemblance). Ce critère peut aussi être utilisé comme un outil de diagnostic. Ainsi, pour un nombre de composants K' donné, si $NEC_{K'} \leq 1$ on peut en conclure qu'il y a bien une structure au sein des données.

Modèles gaussiens parcimonieux :

Dans le cadre des modèles de mélange gaussiens, le nombre de paramètres à estimer peut être important et le nombre de données disponibles insuffisant pour obtenir une estimation fiable. Une méthode classique consiste à réduire le nombre de paramètres en appliquant des contraintes sur la matrice de variance-covarianceΣ_k. Bandfield et Raftery (1993) et Celeux et Govaert (1995) ont proposé une paramétrisation de la matrice Σ_kà partir de sa décomposition en valeurs propres :

$$\Sigma_k = \lambda_k D_k A_k D_k'$$

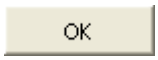
où $\lambda_k = |\Sigma_k|^{\frac{1}{d}}$ est le volume du kième composant, D_k la matrice de vecteurs propres et A_k est une matrice diagonale composée des valeurs propres de Σ_k rangées dans l'ordre décroissant, telle que $|A_k| = 1$. Ces deux matrices D_k et A_k permettent de contrôler respectivement l'orientation et la forme du kième composant.

Modèle	Nombre de paramètres	Nom du modèle
$\lambda_k D_k A_k D_k'$	$a + Kb$	VVV
$\lambda D A D'$	$a + b$	EEE
$\lambda D_k A D_k'$	$a + Kb - (K-1)d$	EEV
$\lambda D_k A_k D_k'$	$a + Kb - (K-1)$	EVV
$\lambda_k D_k A D_k'$	$a + Kb - (K-1)(d-1)$	VEV
λB	$a + d$	EEI
λB_k	$a + Kd - K + 1$	EVI
$\lambda_k B_k$	$a + Kd$	VVI
$\lambda_k B$	$a + d + K - 1$	VEI
λI	$a + 1$	EII
$\lambda_k I$	$a + d$	VII
$\lambda_k D A D$	$a + b + K - 1$	VEE
$\lambda D A_k D$	$a + Kb + (K-1)(d-1)$	EVE
$\lambda_k D A_k D$	$a + Kb + (K-1)d$	VVE

On dispose alors de 28 modèles différents dans le cas multidimensionnel. Concernant l'analyse unidimensionnele, uniquement deux modèles sont envisageable (variance égale ou non).

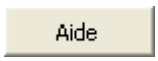
Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.



: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.

Onglet **Général**:

Tableau observations/variables : sélectionnez un tableau comprenant N objets décrits par P descripteurs. Si des en-têtes de colonnes ont été sélectionnés, veuillez vérifier que l'option « Libellés des variables » est activée.

Poids des lignes : activez cette option si vous voulez pondérer les lignes. Si vous n'activez pas cette option, les poids seront tous considérés comme valant 1. Les poids doivent être impérativement supérieurs ou égaux à 0. Si un en-tête de colonne a été sélectionné, veuillez vérifier que l'option « Libellés des colonnes » est activée.

Etiquetage partiel : activez cette option si vous souhaitez spécifier le groupe d'appartenance de certaines observations. Si vous n'activez pas cette option, les groupes d'appartenance seront tous considérés comme inconnus. Les valeurs associées aux groupes doivent être impérativement des entiers supérieurs ou égaux à 1, si le groupe d'appartenance est inconnu, la cellule associée doit rester vide. Si un en-tête de colonne a été sélectionné, veuillez vérifier que l'option « Libellés des colonnes » est activée.

Dimension des données : vous avez la possibilité de réaliser une analyse unidimensionnelle (colonne par colonne) ou multidimensionnelle (toutes les colonnes).

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des colonnes : activez cette option si la première ligne des données sélectionnées (Tableau observations/variables, poids des lignes) contient un libellé.

Libellés des lignes : activez cette option si vous voulez utiliser des libellés d'observations pour l'affichage des résultats. Si l'option « Libellés des colonnes » est activée, la première cellule de la sélection doit comprendre un en-tête. Si vous n'activez pas cette option, des libellés seront automatiquement créés (Obs1, Obs2, …).

Onglet **Options (1)** :

Algorithme d'inférence : sélectionnez l'algorithme avec lequel l'inférence des paramètres sera réalisée.

- *EM :* l'algorithme usuel EM proposé par Dempster et al. est utilisé. Il s'agit de l'algorithme utilisé par défaut.
- *SEM :* version stochastique de l'algorithme EM. Une étape stochastique est ajoutée pour affecter les individus aux différents groupes. Cet algorithme peut mener à des classes vides et perturber l'estimation des paramètres.
- *CEM :* version classifiant de l'algorithme EM. Une étape de classification est ajoutée pour affecter les individus aux groupes par la règle du MAP (Maximum A Posteriori). Cet algorithme peut mener à des classes vides et perturber l'estimation des paramètres.

Critères de sélection : sélectionnez le critère avec lequel le nombre de composants est sélectionné.

- *BIC :* le Bayesian Information Criterion. Il s'agit du critère utilisé par défaut.
- *AIC :* le Akaike Information Criterion. Ce critère a tendance à surestimer le nombre de composants.
- *ICL :* le Integrated Complete Likelihood. Ce critère recherche le modèle qui fournit les groupes les mieux séparés. Généralement, le nombre de composants sélectionné est inférieur à celui obtenu par BIC.
- *NEC :* le Normalized Entropy Criterion. Ce critère recherche la meilleure séparation entre les groupes. Le NEC n'est pas défini pour un modèle avec un composant. Ce critère permet de choisir le nombre de composants et non la paramétrisation du modèle.

Initialisation : sélectionnez la méthode d'initialisation de l'algorithme.

- *Aléatoire :* une partition aléatoire des données est utilisée pour initialiser l'algorithme. L'algorithme sera lancé autant de fois que spécifié par le nombre de répétitions choisi jusqu'à la convergence de l'algorithme. La meilleure estimation parmi toutes les répétitions est retenue.
- *Court EM :* une partition aléatoire des données est utilisée pour initialiser l'algorithme puis l'algorithme sera lancé autant de fois que spécifié par le nombre de répétitions choisi avec un maximum de 5 itérations. La meilleure partition obtenue sera retenue pour initialiser l'algorithme.

- *K-means :* la partition obtenue par l'algorithme des K-means est utilisée pour initialiser l'algorithme.

Nombre de répétitions : spécifiez le nombre de répétitions lorsque la méthode d'initialisation choisie est *Aléatoire* ou *Court EM*.

Conditions d' arrêt :

Itérations : entrez le nombre maximal d'itérations pour l'algorithme d'inférence. Les calculs sont interrompus dès que le nombre maximal d'itérations est dépassé. Valeur par défaut : 500.

Convergence : entrez la valeur minimale d'évolution des paramètres d'une itération à l'autre, qui une fois atteinte permet de considérer que l'algorithme a convergé. Valeur par défaut : 0,00001.

Onglet **Options (2)** :

Modèles de mélange : sélectionnez le/les modèle(s) que vous souhaitez utiliser pour modéliser les données. Le meilleur de ces modèles sera retenu par rapport au critère de sélection choisi.

Nombre de classes : spécifiez le nombre minimal et maximal de classes souhaités. Le nombre minimal doit être supérieur ou égal à 1 et le nombre maximal inférieur au nombre de données. Par défaut, le nombre de classes varie de 2 à 5.

Proportions égales : activez cette option pour contraindre les proportions des groupes à être égales.

Onglet **Données manquantes** :

Ne pas accepter les valeurs manquantes : activez cette option pour que XLSTAT empêche la poursuite des calculs si des valeurs manquantes sont détectées.

Supprimer les observations : activez cette option pour supprimer les observations comportant des données manquantes.

Onglet **Sorties**:

Statistiques descriptives : activez cette option pour afficher les statistiques descriptives pour les variables sélectionnées.

Evolution du critère : activez cette option pour obtenir le tableau des valeurs du critère de sélection pour l'ensemble des modèles choisis.

Probabilités *a posteriori* : activez cette option pour obtenir les probabilités d'appartenance de chaque observation aux différents groupes.

Classification MAP : activez cette option pour obtenir l'affectation des observations aux différents groupes par la règle du MAP.

Onglet **Graphiques** :

Evolution du critère : activez cette option pour obtenir le graphique représentant l'évolution du critère de sélection pour l'ensemble des modèles choisis.

Classification MAP : activez cette option pour obtenir le graphique des données classées selon la règle du MAP. Chaque couleur correspond à un groupe différent.

Modèle ajusté : activez cette option pour obtenir la représentation du modèle sélectionné.

Fonction de répartition : activez cette option pour obtenir la représentation de la fonction de répartition empirique et estimée. Ce graphique est un outil de diagnostic, plus les fonctions de répartition sont proches et plus l'ajustement est bon. Ce graphique est disponible uniquement dans le cas unidimensionnel.

Q-Q plot : activez cette option pour obtenir le Q-Q plot entre la distribution empirique et celle estimée. Ce graphique est un outil de diagnostic, plus la courbe obtenue est proche de la première bissectrice et plus l'ajustement est bon. Ce graphique est disponible uniquement dans le cas unidimensionnel.

Résultats

Statistiques descriptives : dans ce tableau sont affichées les statistiques descriptives correspondant aux différents éléments.

Evolution du critère : dans ce tableau sont affichées les valeurs du critère de sélection pour l'ensemble des modèles choisis. Ces valeurs sont également représentées sous forme de graphique.

Paramètres estimés : les proportions, moyennes et variances par classes sont fournies pour le modèle sélectionné.

Caractéristiques du modèle sélectionné : dans ce tableau, plusieurs caractéristiques du modèle sont indiquées (BIC, AIC, ICL, Log-vraisemblance, NEC, Entropie, DDL).

Probabilités a posteriori : dans ce tableau sont affichées les probabilités d'appartenance de chaque observation aux différents groupes.

Classification MAP : dans ce tableau est indiquée l'affectation des observations aux différents groupes par la règle du MAP. Cette classification est aussi représentée sous forme de graphique.

Modèle ajusté : ce graphique représente l'ajustement du modèle sélectionné.

Fonction de répartition : ce graphique permet de comparer visuellement la fonction de répartition empirique à celle estimée.

Q-Q plot : ce graphique permet de comparer visuellement les quantiles de la distribution empirique à celle estimée.

Exemple

Un exemple d'application du modèle de mélange gaussien est disponible sur le Centre d'aide XLSTAT :

http://www.xlstat.com/demo-gmmf.htm

Bibliographie

Akaike H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, **19 (6)** : 716-723.

Banfield J. D. and Raftery A. E. (1993), Model-based gaussian and non- gaussian clustering. *Biometrics*, **49**, 803-821.

Biernacki C., Celeux G. and Govaert G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **22**, 719-725.

Celeux G. and Govaert G. (1992). A classification EM algorithm for clustering and two stochastic versions. *Computational Statistics & Data Analysis*, **14**, 315-332.

Celeux G. and Govaert G. (1995). Parsimonious Gaussian models in cluster analysis. *Pattern Recognition*, **28**, 781-793.

Celeux G. and Soromenho G. (1996). An entropy criterion for assessing the number of clusters in a mixture model. *Journal of Classification*, **13**, 195-212.

Dempster A. P., Laird N. M. and Rubin D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *JRSS*, **39**, 1-38.

McLachlan, G. J. and Peel D. (2000). Finite Mixture Models. New York, Wiley.

Schwarz G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2), 461-464.

Partitionnement univarié

Utilisez le partitionnement univarié pour regrouper de façon optimale des objets dans k classes homogènes, sur la base de leur description par une seule variable quantitative.

Dans cette section :

Description

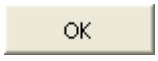
Le partitionnement univarié consiste à regrouper N observations unidimensionnelles (décrites par une seule variable quantitative) dans k classes homogènes.

L'homogénéité est ici mesurée au travers de la somme des variances intra- classe. Pour maximiser l'homogénéité des classes, on cherche donc à minimiser la somme des variances intra-classe.

L'algorithme utilisé ici, très rapide, s'appuie sur la méthode proposée par W.D. Fisher (1958). Cette méthode peut être vue comme une discrétisation d'une variable quantitative en une variable ordinale. Les applications sont très nombreuses, avec par exemple des applications en cartographie pour la création d'échelles de couleur ou en marketing pour la création de segments homogènes.

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.

 : cliquez sur ce bouton pour effacer les sélections de données.

Onglet **Général**:

Tableau observations/variables : sélectionnez un tableau comprenant N objets décrits par P descripteurs. Si des en-têtes de colonnes ont été sélectionnés, veuillez vérifier que l'option « Libellés des variables » est activée. Si plusieurs variables sont sélectionnées, elles seront chacune à leur tour partitionnées.

Poids des lignes : activez cette option si vous voulez pondérer les lignes. Si vous n'activez pas cette option, les poids seront tous considérés comme valant 1. Les poids doivent être impérativement supérieurs ou égaux à 0. Si un en-tête de colonne a été sélectionné, veuillez vérifier que l'option « Libellés des colonnes » est activée.

Nombre de classes : entrez le nombre de classes qui doivent être créées par l'algorithme.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des colonnes : activez cette option si la première ligne des données sélectionnées (Tableau observations/variables, libellés des lignes, poids des lignes, poids des colonnes) contient un libellé.

Libellés des lignes : activez cette option si vous voulez utiliser des libellés d'observations pour l'affichage des résultats. Si l'option « Libellés des colonnes » est activée, la première cellule de la sélection doit comprendre un en-tête. Si vous n'activez pas cette option, des libellés seront automatiquement créés (Obs1, Obs2, …).

Onglet **Données manquantes** :

Ne pas accepter les valeurs manquantes : activez cette option pour que XLSTAT empêche la poursuite des calculs si des valeurs manquantes sont détectées.

Supprimer les observations : activez cette option pour supprimer les observations comportant des données manquantes.

Estimer les données manquantes : activez cette option pour estimer les données manquantes avant le début des calculs.

- **Moyenne ou mode** : activez cette option pour estimer les données manquantes en utilisant la moyenne (variables quantitatives) ou le mode (variables qualitatives) pour les variables correspondantes.
- **Plus proche voisin** : activez cette option pour estimer les données manquantes d'une observation en recherchant le plus proche voisin de l'observation.

Onglet **Sorties**:

Statistiques descriptives : activez cette option pour afficher les statistiques descriptives pour les variables sélectionnées.

Barycentres : activez cette option pour afficher les coordonnées des barycentres des classes.

Objets centraux : activez cette option pour afficher les coordonnées de l'objet le plus proche du barycentre de chaque classe.

Résultats par classe : activez cette option pour afficher un tableau donnant les statistiques et les objets correspondant à chacune des classes.

Résultats par objet : activez cette option pour afficher un tableau donnant pour chaque objet sa classe d'affectation dans l'ordre initial des objets.

Résultats

Statistiques simples : dans ce tableau sont affichés pour les variables sélectionnées, le nombre d'observations, le nombre de données manquantes, le nombre de données non manquantes, la moyenne, et l'écart-type.

Barycentres des classes : dans ce tableau sont affichées les coordonnées des barycentres des classes pour les différents descripteurs.

Distances entre les barycentres des classes : dans ce tableau sont affichées les distances euclidiennes entre les barycentres des classes pour les différents descripteurs.

Objets centraux : dans ce tableau sont affichées pour chaque classe les coordonnées de l'objet le plus proche du barycentre de la classe.

Distances entre les objets centraux : dans ce tableau sont affichées les distances euclidiennes entre les objets centraux des classes pour les différents descripteurs.

Résultats par classe : les statistiques descriptives des classes (nombre d'objets, somme des poids, variance intra-classe, distance minimale au barycentre, distance maximale au barycentre, distance moyenne au barycentre) sont affichées dans la première partie du tableau. Dans la seconde partie sont affichés les objets.

Résultats par objet : dans ce tableau est indiquée, pour chaque objet, sa classe d'affectation dans l'ordre initial des objets.

Exemple

http://www.xlstat.com/demo-UniClusterf.htm

Bibliographie

Fisher W.D. (1958). On grouping for maximum homogeneity. *Journal of the American Statistical Association*, **53**, 789-798.

Modélisation des données

Ajustement d'une loi de probabilité

Utilisez ce module pour ajuster une loi de probabilité à un échantillon de données quantitatives continues ou discrètes.

Dans cette section :

Description

L'ajustement d'une loi de distribution à un échantillon de données consiste, une fois le type de loi choisi, à estimer les paramètres de la loi de telle sorte que l'échantillon soit le plus vraisemblable possible (au sens du maximum de vraisemblance) ou qu'au moins certaines statistiques de l'échantillon (moyenne, variance par exemple) correspondent le mieux possible à celles de la loi.

Lois de distribution

XLSTAT permet l'utilisation des lois suivantes :

- Arcsinus (a) : la densité de cette loi (dérivée de la loi Bêta de type I) est donnée par :

$$f(x) = \frac{\sin(\pi\alpha)}{\pi x}(\frac{x}{1-x})^{\alpha-1}, \quad \text{avec } 0 < \alpha < 1, x \in [0,1]$$

On a $E(X) = \alpha$ et $V(X) = \alpha(1-\alpha)/2$

- Bernoulli (p) : la densité de cette loi est donnée par :

$$P(X=1) = p, P(X=0) = 1-p, \quad \text{avec } p \in [0,1]$$

On a $E(X) = p$ et $V(X) = p(1-p)$

La loi de Bernoulli, du nom du mathématicien suisse Jacob Bernoulli (1654-1705), permet de décrire les phénomènes aléatoires binaires où seuls deux événements peuvent survenir avec

des probabilités respectives de p et $1-p$.

- Bêta (a, b) : la densité de cette loi (aussi appelée Bêta de type I) est donnée par :

$$f(x) = \frac{1}{B(\alpha, \beta)} x^{\alpha-1}(1-x)^{\beta-1}$$

$$\text{avec} \quad \alpha, \beta > 0, x \in [0,1] \text{ et } B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

On a $E(X) = \alpha/(\alpha+\beta)$ et $V(X) = \alpha\beta/[(\alpha+\beta+1)(\alpha+\beta)^2]$

- Bêta4 (α, β, c, d) : la densité de cette loi est donnée par :

$$f(x) = \frac{1}{B(\alpha, \beta)} \frac{(x-c)^{\alpha-1}(d-x)^{\beta-1}}{(d-c)^{\alpha+\beta-1}}, \quad \text{avec } \alpha, \beta > 0, x \in [c, d]$$

$$c, d \in \mathbb{R} \text{ et } B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

On a $E(X) = \frac{c+(c-d)\alpha}{(\alpha+\beta)}$ et $V(X) = \frac{(c-d)^2\alpha\beta}{(\alpha+\beta+1)(\alpha+\beta)^2}$

Pour la loi Bêta de type I, la distribution est dans l'intervalle $[0, 1]$. La loi Bêta4 est obtenue par un simple changement de variable de la loi Bêta de type I de telle sorte que la distribution soit sur l'intervalle [c, d].

- Binomiale (n, p) : la densité de cette loi est donnée par :

$$P(X = x) = C_n^x p^x (1-p)^{n-x}, \quad \text{avec } n, x \in \mathbb{N}, n > 0, x \in [0, n], p \in [0, 1]$$

On a $E(X) = np$ et $V(X) = np(1-p)$

n est le nombre d'essais, et p la probabilité de succès. La loi binomiale est la loi du nombre de succès pour n essais, sachant que la probabilité de succès vaut p. La loi binomiale peut être vue comme la loi de n tirages dans une loi de Bernoulli.

- Binomiale négative (n, p) de type I : la densité de cette loi est donnée par :

$$P(X = x) = C_{n+x-1}^{x-1} p^n (1-p)^x, \quad \text{avec } n, x \in \mathbb{N}, n > 0, p \in [0, 1]$$

On a $E(X) = n(1-p)/p$ et $V(X) = n(1-p)/p^2$

n est le nombre de succès et p la probabilité de succès. La loi binomiale négative de type I est la loi du nombre de tirages x sans succès nécessaires avant d'avoir obtenus n succès.

- Binomiale négative (k, p) de type II : la densité de cette loi est donnée par :

$$P(X = x) = \frac{\Gamma(k+x)p^x}{x!\Gamma(k)(1+p)^{k+x}}, \quad \text{avec } x \in N, k, p > 0$$

On a $E(X) = kp$ et $V(X) = kp(p+1)$

La loi binomiale négative de type II permet de représenter des phénomènes discrets fortement hétérogènes. Lorsque k tend vers l'infini, la loi binomiale négative de type II tend vers une loi de Poisson de paramètre $\lambda = kp$.

- $Khi^2(df)$: la densité de cette loi est donnée par :

$$f(x) = \frac{(1/2)^{df/2}}{\Gamma(df/2)} x^{\frac{df}{2}-1} e^{-x/2}, \quad \text{avec } x > 0, df \in \mathbb{N}^*$$

On a $E(X) = df$ et $V(X) = 2df$

La loi du Khi² correspond à la loi de la somme des carrés de df lois normales centrées réduites (lois normales standard). Elle est très utilisée pour tester des hypothèses.

- Erlang (k, λ) : la densité de cette loi est donnée par :

$$f(x) = \lambda^k x^{k-1} \frac{e^{-\lambda x}}{(k-1)!}, \quad \text{avec } x \geq 0 \text{ et } k, \lambda > 0 \text{ et } k \in \mathbb{N}$$

On a $E(X) = k/\lambda$ et $V(X) = k/\lambda^2$

k est le paramètre de forme de la loi et λ est le paramètre de taux.

Cette distribution, développée par le scientifique danois A. K. Erlang (1878-1929) pour l'étude du trafic téléphonique, est utilisée de manière plus générale pour l'étude des files d'attente.

Remarque : lorsque $k = 1$, cette distribution est équivalente à la distribution exponentielle, et la loi Gamma à deux paramètres est une généralisation de la loi d'Erlang au cas où k est un réel et non un entier (par ailleurs on utilise le paramètre d'échelle $\beta = 1/\lambda$).

- Exponentielle (λ) : la densité de cette loi est donnée par :

$$f(x) = \lambda \exp(-\lambda x), \quad \text{avec } x > 0 \text{ et } \lambda > 0$$

On a $E(X) = 1/\lambda$ et $V(X) = 1/\lambda^2$

La loi exponentielle est souvent utilisée pour étudier la durée de vie en contrôle qualité.

- Fisher (df_1, df_2) : la densité de cette loi est donnée par :

$$f(x) = \frac{1}{xB(df_1/2, df_2/2)} \left(\frac{df_1 x}{df_1 x + df_2}\right)^{df_1/2} \left(1 - \frac{df_1 x}{df_1 x + df_2}\right)^{df_2/2}$$

$$\text{avec } x > 0 \text{ et } df_1, df_2 \in \mathbb{N}^*$$

On a $E(X) = df_2/(df_2 - 2)$ si $df_2 > 2$, et $V(X) = \frac{2df_2^2(df_1+df_2-2)}{df_1(df_2-2)^2(df_2-4)}$ si $df_2 > 4$

La loi de Fisher, du nom du biologiste, généticien et statisticien Ronald Aylmer Fisher (1890-1962), correspond au rapport de deux lois du Khi². Elle est très utilisée pour tester des hypothèses.

- Fisher-Tippett (β, μ) : la densité de cette loi est donnée par :

$$f(x) = \frac{1}{\beta} \exp\left(-\frac{x-\mu}{\beta} - \exp\left(-\frac{x-\mu}{\beta}\right)\right), \quad \text{avec } \beta > 0$$

On a $E(X) = \mu + \beta\gamma$ et $V(X) = (\pi\beta)^2/6$ où γ est la constante de Euler-Mascheroni.

La loi de Fisher-Tippett, aussi appelée loi Log-Weibull, ou loi généralisée des valeurs extrêmes, est utilisée dans l'étude de phénomènes extrêmes. La loi de Gumbel est un cas particulier de la loi de Fisher-Tippett avec $\beta = 1$ et $\mu = 0$.

- Gamma (k, β, μ) : la densité de cette loi est donnée par :

$$f(x) = (x-\mu)^{k-1} \frac{e^{-(x-\mu)/\beta}}{\beta^k \Gamma(k)}, \quad \text{avec } x > \mu \text{ et } k, \beta > 0$$

On a $E(X) = \mu + k\beta$ et $V(X) = k\beta^2$

k est le paramètre de forme de la loi et β est le paramètre d'échelle.

- GEV (β, k, μ): la densité de cette loi est donnée par :

$$f(x) = \frac{1}{\beta}\left(1 + k\frac{x-\mu}{\beta}\right)^{-1/k-1} \exp\left(-\left(1 + k\frac{x-\mu}{\beta}\right)^{-1/k}\right), \quad \text{avec} \quad \beta > 0$$

On a $E(X) = \mu + \frac{\beta}{k}\Gamma(1+k)$ et $V(X) = \left(\frac{\beta}{k}\right)^2 \left(\Gamma(1+2k) - \Gamma^2(1+k)\right)$

La loi GEV (Generalized Extreme Values) est très utilisée en hydrologie pour modéliser les phénomènes de crues. k est classiquement compris entre -0.6 et 0.6.

- Gumbel : la densité de cette loi est donnée par :

$$f(x) = \exp(-x - \exp(-x))$$

On a $E(X) = \gamma$ et $V(X) = \pi^2/6$ où γ est la constante de Euler-Mascheroni (0.5772156649…).

La loi de Gumbel, du nom de Emil Julius Gumbel (1891-1966), est un cas particulier de la loi de Fisher-Tippett avec $\beta = 1$ et $\mu = 0$. Elle est utilisée dans l'étude de phénomènes extrêmes comme les précipitations ou les crues maximales et les magnitudes maximales de tremblement de terre.

- Logistique (μ, s) : la densité de cette loi est donnée par :

$$f(x) = \frac{e^{-\frac{(x-\mu)}{s}}}{s(1+e^{-\frac{(x-\mu)}{s}})}, \quad \text{avec } s > 0$$

On a $E(X) = \mu$ et $V(X) = (\pi s)^2/3$

- Lognormale (μ, σ) : la densité de cette loi est donnée par :

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(\ln(x)-\mu)^2}{2\sigma^2}}, \quad \text{avec } x, \sigma > 0$$

On a $E(X) = exp(\mu + \sigma^2/2)$ et $V(X) = \left(exp(\sigma^2) - 1\right) exp(2\mu + \sigma^2)$

- Lognormale2 (m, s) : la densité de cette loi est donnée par :

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(\ln(x)-\mu)^2}{2\sigma^2}}, \text{ avec } x, \sigma > 0$$

On a :

$\mu = ln(m) - ln(1 + s^2/m^2)/2$ et $\sigma^2 = ln(1 + s^2/m^2)$

Et :

$E(X) = m$ et $V(X) = s^2$

Cette distribution est simplement une reparamétrisation de la loi Lognormale.

- Normale (μ, σ) : la densité de cette loi est donnée par :

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad \text{avec } \sigma > 0$$

On a $E(X) = \mu$ et $V(X) = \sigma^2$

- Normale standard : la densité de cette loi est donnée par :

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

On a $E(X) = 0$ et $V(X) = 1$

Cette loi est un cas particulier de la loi normale, avec $\mu = 0$ et $\sigma = 1$. Elle est aussi appelée loi normale centrée réduite.

- Pareto (a, b) : la densité de cette loi est donnée par :

$$f(x) = \frac{ab^a}{x^{a+1}}, \text{ avec } a, b > 0 \text{ et } x \geq b$$

On a $E(X) = ab/(a-1)$ et $V(X) = \frac{ab^2}{((a-1)^2(a-2))}$

La loi de Pareto, du nom de l'économiste italien Vilfredo Pareto (1848-1923), est aussi connue sous le nom de loi de Bradford. Cette loi a d'abord été utilisée pour représenter la répartition des richesses dans la société, avec notamment le principe de Pareto, selon lequel 80% des richesses d'un pays sont détenus par 20% de la population.

- PERT (a, m, b) : la densité de cette loi est donnée par :

$$f(x) = \frac{1}{B(\alpha, \beta)} \frac{(x-a)^{\alpha-1}(b-x)^{\beta-1}}{(b-a)^{\alpha+\beta-1}}, \text{ avec } \alpha, \beta > 0, x \in [a, b]$$

$$a, b \in \mathbb{R} \text{ et } B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

$$\alpha = \frac{4m + b - 5a}{b-a}$$

$$\beta = \frac{5b - a - 4m}{b-a}$$

On a $E(X) = (b-a)\alpha/(\alpha+\beta)$ et $V(X) = (b-a)^2\alpha\beta/\left((\alpha+\beta+1)\left(\alpha+\beta\right)^2\right)$

La loi de PERT est donc un cas particulier de la loi Bêta4, définie par son intervalle de définition $[a, b]$ et sa valeur la plus probable m (le mode). PERT est l'acronyme de *Program Evaluation and Review Technique* , une méthode de gestion et de planification de projet. La méthodologie et la distribution PERT ont été utilisées pour la première fois pour le projet de développement des missiles Polaris lancés depuis des sous-marins par la marine américaine et Lockheed de 1956 à 1960 (Clark 1962). La distribution PERT permet de modéliser le temps probable nécessaire à une équipe pour terminer son projet. La loi triangulaire, plus simple, permet aussi de modéliser ce type de phénomènes avec les trois mêmes paramètres.

- Poisson (λ): la densité de cette loi est donnée par :

$$P(X = x) = \frac{\exp(-\lambda)\lambda^x}{x!}, \quad \text{avec } x \in \mathbb{N} \text{ et } \lambda > 0$$

On a $E(X) = \lambda$ et $V(X) = \lambda$

La loi de Poisson, découverte par le mathématicien et astronome Siméon-Denis Poisson (1781-1840) qui fut élève de Laplace, Lagrange et Legendre, est souvent utilisée pour étudier des phénomènes de file d'attente.

- Student (df) : la densité de cette loi est donnée par :

$$f(x) = \frac{\Gamma((df+1/2))}{\sqrt{\pi df}\Gamma(df/2)}(1 + x^2/df)^{-(df+1)/2}, \text{ avec } df > 0$$

On a $E(X) = 0$ si $df > 1$ et $V(X) = df/(df-2)$ si $df > 2$

La loi de Student, du nom que se donnait le chimiste et statisticien anglais William Sealy Gosset (1876-1937) afin de préserver son anonymat (la brasserie Guinness interdisait à ses employés de publier, suite à la publication par un autre chercheur d'informations confidentielles) est la loi de la moyenne de df variables distribuées suivant une loi normale centrée réduite. Lorsque $df = 1$, la loi de Student est une loi de Cauchy dont la particularité est de n'avoir ni espérance ni variance.

- Trapézoïdale (a, b, c, d) : la densité de cette loi est donnée par :

$$\begin{cases} f(x) = \frac{2(x-a)}{(d+c-b-a)(b-a)}, & x \in [a, b] \\ f(x) = \frac{2}{(d+c-b-a)}, & x \in [b, c] \\ f(x) = \frac{2(d-x)}{(d+c-b-a)(d-c)}, & x \in [c, d] \\ f(x) = 0, & x < a, \quad x > d \\ \text{avec } a < b < c < d \end{cases}$$

On a $E(X) = \frac{d^2+c^2-b^2-a^2+cd-ab}{3(d+c-b-a)}$ et $V(X) = \frac{(c+d)(c^2+d^2)-(a+b)(a^2+b^2)}{6(d+c-b-a)} - E^2(X)$

Cette loi est utile pour représenter un phénomène dont on sait qu'il peut prendre des valeurs entre deux extrêmes, mais pour lequel un intervalle plus restreint paraît plus raisonnable.

- Triangulaire (a, m, b) : la densité de cette loi est donnée par :

$$\begin{cases} f(x) = \frac{2(x-a)}{(b-a)(m-a)}, & x \in [a, m] \\ f(x) = \frac{2(b-x)}{(b-a)(b-m)}, & x \in [m, b] \\ f(x) = 0, & x < a, \quad x < b \\ \text{avec } a < m < b \end{cases}$$

On a $E(X) = (a + m + b)/3$ et $V(X) = (a^2 + m^2 + b^2 - ab - am - bm)/18$

- TriangulaireQ (q_1, m, q_2, p_1, p_2): cette loi est une reparamétrisation de la loi triangulaire. Une première étape nécessite l'estimation des paramètres a et b de la distribution triangulaire pour savoir à quels quantiles q_1 et q_2 correspondent les pourcentages p_1 et p_2. Une fois ceci fait, on peut utiliser la fonction de densité ou de répartition triangulaire.

- Uniforme (a, b) : la densité de cette loi est donnée par :

$$f(x) = \frac{1}{b-a}, \text{ avec } b > a \text{ et } x \in [a,b]$$

On a $E(X) = (a+b)/2$ et $V(X) = (b-a)^2/12$

La loi uniforme (0, 1) est très utilisée pour les simulations. Comme la fonction de répartition de toutes les lois est comprise entre 0 et 1, un échantillon tiré dans une loi Uniforme (0,1) permet d'obtenir un échantillon dans toutes les lois dont on sait calculer l'inverse.

- Uniforme discrète (a,b) : la densité de cette loi est donnée par :

$$P[X = x] = \frac{1}{b-a+1}, \text{ avec } (a,b,x) \in \mathbb{N}^3, x \in [a,b]$$

On a $E(X) = (a+b)/2$ et $V(X) = [(b-a+1)^2 - 1]/12$

La loi uniforme discrète correspond au cas particulier où la loi uniforme est restreinte à des nombre entiers.

- Weibull (β) : la densité de cette loi est donnée par :

$$f(x) = \beta x^{\beta-1} \exp(-x^\beta), \text{ avec } x > 0 \text{ et } \beta > 0$$

On a $E(X) = \Gamma(\frac{1}{\beta} + 1)$ et $V(X) = \Gamma(\frac{2}{\beta} + 1) - \Gamma^2(\frac{1}{\beta} + 1)$

Le paramètre β est le paramètre de forme de la loi de Weibull.

- Weibull (β, γ) : la densité de cette loi est donnée par :

$$f(x) = \frac{\beta}{\gamma}(\frac{x}{\gamma})^{\beta-1} e^{-(\frac{x}{\gamma})^\beta}, \text{ avec } x > 0, \text{ et } \beta, \gamma > 0$$

On a $E(X) = \gamma\Gamma\left(\frac{1}{\beta} + 1\right)$ et $V(X) = \gamma^2\left[\Gamma\left(\frac{2}{\beta} + 1\right) - \Gamma^2\left(\frac{1}{\beta} + 1\right)\right]$

Le paramètre β est le paramètre de forme et le paramètre γ est le paramètre d'échelle. Lorsque $\beta = 1$, la loi de Weibull est une loi exponentielle de paramètre $1/\gamma$.

- Weibull (β, γ, μ): la densité de cette loi est donnée par :

$$f(x) = \frac{\beta}{\gamma}\left(\frac{x-\mu}{\gamma}\right)^{\beta-1} e^{-\left(\frac{x-\mu}{\gamma}\right)^\beta}, \text{ avec } x > \mu, \text{ et } \beta, \gamma > 0$$

On a $E(X) = \mu + \gamma\Gamma\left(\frac{1}{\beta} + 1\right)$ et $V(X) = \gamma^2\left[\Gamma\left(\frac{2}{\beta} + 1\right) - \Gamma^2\left(\frac{1}{\beta} + 1\right)\right]$

La loi de Weibull, du nom du suédois Ernst Hjalmar Waloddi Weibull (1887-1979), est très utilisée en contrôle qualité et en analyse de survie. Le paramètre β est le paramètre de forme et le paramètre γ est le paramètre d'échelle. Lorsque $\beta = 1$ et $\mu = 0$, la loi de Weibull est une loi exponentielle de paramètre $1/\gamma$.

Méthodes d'ajustement

Deux méthodes d'ajustement sont proposées par XLSTAT :

Moments : cette méthode simple utilise la définition des moments de la loi en fonction des paramètres afin de déterminer ces derniers. Pour la plupart des lois, l'utilisation de la moyenne et de la variance est suffisante. Cependant, pour certaines lois la moyenne suffit (par exemple, la loi de Poisson), ou, au contraire, le coefficient d'asymétrie est aussi nécessaire (loi de Weibull par exemple).

Vraisemblance : les paramètres de la loi sont estimés en maximisant la vraisemblance de l'échantillon. Cette méthode, plus complexe, présente l'avantage d'être rigoureuse pour toutes les lois, et de permettre d'obtenir des écart-types approximatifs pour les estimateurs des paramètres. La méthode du maximum de vraisemblance est proposée pour la loi binomiale négative de type II, la loi de Fisher-Tippett, la loi GEV et la loi de Weibull.

Pour certaines lois, la méthode des moments donne exactement le même résultat que celle du maximum de vraisemblance. C'est notamment le cas pour la loi normale.

Tests d'ajustement

Une fois que les paramètres de la loi choisie sont déterminés, pour vérifier si le phénomène observé au travers de l'échantillon suit la loi en question, il est nécessaire de tester l'hypothèse. Deux tests d'ajustement sont proposés par XLSTAT.

Le **test d'ajustement du** Khi^2 est un test paramétrique utilisant la distance (au sens du Khi^2) entre l'histogramme de la distribution théorique (déterminée par les paramètres estimés) et l'histogramme de la distribution empirique de l'échantillon. Les histogrammes sont calculés en utilisant k intervalles choisis par l'utilisateur. On montre que la statistique calculée suit asymptotiquement une loi du Khi^2 à $(n-k)$ degrés de liberté, où n est l'effectif de l'échantillon. Ce test est plutôt recommandé pour les distributions discrètes, et il est conseillé de veiller à ce que l'espérance de l'effectif de chacune des classes ne soit pas inférieur à 5.

Il peut arriver que le test du Khi^2 amène à conclure à un mauvais ajustement de la distribution aux données avec une classe contribuant beaucoup plus au Khi^2 que les autres. Dans un tel cas, la réunion de la classe en question avec une classe voisine permet de vérifier si la conclusion est uniquement due à la classe en question, ou si l'ajustement est réellement mauvais.

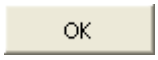
Le **test d'ajustement deKolmogorov-Smirnov** est un test non paramétrique exact basé sur la distance maximale entre une fonction de répartition théorique (entièrement déterminée par les valeurs connues de ses paramètres) et la fonction de répartition empirique de l'échantillon. Ce test n'est utilisable que pour les distributions continues.

Dans le cas où une estimation des paramètres précède le test d'ajustement, le test de Kolmogorov-Smirnov n'est pas correct, puisque les paramètres sont estimés en essayant de rapprocher la distribution théorique le plus possible des données. Le test de Kolmogorov-Smirnov, s'il valide l'hypothèse de bon ajustement, risque d'être optimiste.

Pour le cas où la loi utilisée est la loi normale, Lilliefors et Stephens (voir tests de normalité) ont proposé un test de Kolmogorov-Smirnov modifié qui permet l'estimation des paramètres sur l'échantillon testé.

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT vous permet de sélectionner les données par colonnes ou par plage. Si la flèche est vers la droite, XLSTAT vous permet de sélectionner les données par lignes ou par plage.

Onglet **Général**:

Données : sélectionnez les données correspondant à l'échantillon de données pour lequel le test d'ajustement doit être calculé. Vous pouvez sélectionner plusieurs colonnes (mode colonnes) ou lignes (mode lignes) si vous voulez effectuer les tests sur plusieurs échantillons en une seule fois.

Distribution : choisissez la loi de probabilité qui doit être utilisée pour l'ajustement et/ou les tests d'ajustement. Voir la partie description pour plus d'information sur les lois proposées. L'option **automatique** permet de laisser XLSTAT identifier la distribution s'ajuste le mieux (déterminé sur la base d'un test de Kolmogorov-Smirnov).

Paramètres : vous pouvez choisir d'**entrer** les paramètres de la loi, ou de les **estimer** . Si vous choisissez d'entrer les paramètres, vous devez entrer la valeur des paramètres.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des échantillons : activez cette option si les libellés des échantillons sont sur la première ligne (mode colonnes) ou dans la première colonne (mode lignes) des données sélectionnées.

Poids : activez cette option si vous voulez pondérer les observations. Si vous n'activez pas cette option, les poids seront tous considérés comme valant 1. Les poids doivent être impérativement supérieurs ou égaux à 0. Si un en- tête de colonne a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

- **Standardiser les poids** : si vous activez cette option, les poids sont standardisés de telle sorte que leur somme soit égale au nombre d'observations.

Onglet **Options**:

Tests : choisissez le type de tests d'ajustement (voir la section description pour plus de détails sur les tests).

- **Kolmogorov-Smirnov** : activez cette option pour effectuer un test de Kolmogorov-Smirnov.
- Khi^2 : activez cette option pour effectuer un test du Khi^2.
- **Niveau de signification (%)** : entrez le niveau de signification pour les tests ci-dessus.

Méthode d'estimation : choisissez la méthode d'estimation des paramètres de la distribution choisie (voir la section description pour plus de détails sur les méthodes d'estimation)

- **Moments** : activez cette option pour utiliser la méthode des moments.
- **Maximum de vraisemblance** : activez cette option pour utiliser la méthode du maximum de vraisemblance. Vous pouvez alors modifier la valeur limite de **convergence** qui, une fois atteinte, permet de considérer que l'algorithme a convergé. Valeur par défaut : 0,00001.

Intervalles : dans le cas d'un test du Khi^2 ou si vous souhaitez comparer la fonction de densité de la loi choisie à l'histogramme de l'échantillon, vous devez choisir l'une des options suivantes :

- **Nombre** : choisissez cette option pour entrer le nombre d'intervalles à créer.
- **Amplitude** : choisissez cette option pour définir une amplitude fixe pour les intervalles.

- **Définis par l'utilisateur** : sélectionnez une colonne contenant en ordre croissant la borne inférieure du premier intervalle, et la borne supérieure de tous les intervalles.
- **Minimum** : activez cette option pour entrer la valeur de la borne inférieure du premier intervalle. Cette valeur doit être inférieure ou égale au minimum de la série.

Onglet **Données manquantes** :

Supprimer les observations :

- **Pour l'échantillon correspondant** : activez cette option pour ne pas prendre en compte une observation dont l'une des données est manquante, uniquement pour les échantillons pour lesquels une donnée est manquante.
- **Pour tous les échantillons** : activez cette option pour ne pas prendre en compte une observation dont l'une des données est manquante, pour tous les échantillons sélectionnés.

Estimer les données manquantes : activez cette option pour estimer les données manquantes en utilisant la moyenne de l'échantillon.

Onglet **Sorties**:

Statistiques descriptives : activez cette option pour afficher les statistiques descriptives pour les échantillons sélectionnés.

Onglet **Graphiques** :

Histogrammes : activez cette option pour afficher les histogrammes des échantillons. Pour la distribution théorique, la fonction de densité est affichée.

- **Barres** : choisissez cette option pour afficher des histogrammes avec une barre pour chaque intervalle.
- **Lignes continues** : choisissez cette option pour afficher des histogrammes avec une ligne continue.

Fonction de répartition empirique : activez cette option pour afficher les histogrammes cumulés des échantillons. Pour la distribution théorique, la fonction de répartition est affichée.

- **Basés sur l'histogramme** : choisissez cette option pour afficher des histogrammes cumulés basés sur la même définition d'intervalles que les histogrammes.
- **Fonction de répartition empirique** : choisissez cette option pour afficher des histogrammes cumulés qui correspondent en réalité à la fonction de répartition empirique de l'échantillon.

Résultats

Statistiques descriptives : dans le tableau des statistiques descriptives sont affichés pour tous les échantillons sélectionnés, le nombre d'observations, le nombre de données manquantes, le nombre de données non manquantes, la moyenne, et l'écart-type.

Paramètres estimés : dans ce tableau sont affichés les paramètres de la loi.

Statistiques estimées à partir des données et calculées à partir des estimateurs des paramètres de la loi : ce tableau permet de comparer la moyenne, la variance, le coefficient d'asymétrie et le coefficient d'aplatissement calculés à partir de l'échantillon à ceux calculés à partir des valeurs des paramètres de la loi.

Test de Kolmogorov-Smirnov : les résultats du test de Kolmogorov-Smirnov sont affichés si l'option correspondante a été activée.

Test du Khi^2 : les résultats du test du Khi^2 sont affichés si l'option correspondante a été activée.

Comparaison entre les effectifs observés et théoriques : ce tableau est affiché si un test du Khi^2 a été demandé.

Statistiques descriptives pour les intervalles : ce tableau est affiché si des histogrammes ont été demandés. Il permet de visualiser les effectifs et les fréquences pour chaque intervalle, ainsi que les densités pour l'échantillon et la distribution choisie.

Exemple

Un exemple d'ajustement d'une loi de probabilité est disponible sur le Centre d'aide XLSTAT :

http://www.xlstat.com/demo-dfitf.htm

Bibliographie

Abramowitz M. and Stegun I.A. (1972) . Handbook of Mathematical Functions. Dover Publications, New York.

Clark C. E. (1962) . The PERT model for the distribution of an activity time. *Operation Research* , **10** (3), 405-406.

El-Shaarawi A.H., Esterby E.S. and Dutka B.J (1981) . Bacterial density in water determined by Poisson or negative binomial distributions. *Applied an Environmental Microbiology* , **41** (1). 107-116.

Fisher R.A. and Tippett H.C. (1928) . Limiting forms of the frequency distribution of the smallest and largest member of a sample. *Proc. Cambridge Phil. Soc.* , **24** , 180-190.

Gumbel E.J. (1941) . Probability interpretation of the observed return periods of floods. *Trans. Am. Geophys. Union* , **21** , 836-850.

Jenkinson A. F. (1955) . The frequency distribution of the annual maximum (or minimum) of meteorological elements. *Q. J. R. Meteorol. Soc.* , **81** , 158-171.

Perreault L. and Bobée B. (1992) . Loi généralisée des valeurs extrêmes. Propriétés mathématiques et statistiques. Estimation des paramètres et des quantiles XT de période de retour T. INRS-Eau, rapport de recherche no 350, Québec.

Weibull W. (1939) . A statistical theory of the strength of material. *Proc. Roy. Swedish Inst. Eng. Res.***151** (1), 1-45.

Régression linéaire

Utilisez ce module pour créer un modèle de régression linéaire simple ou multiple dans un but explicatif ou prédictif.

Dans cette section :

Description

La régression linéaire est sans aucun doute la méthode statistique la plus utilisée. On distingue habituellement la régression simple (une seule variable explicative) de la régression multiple (plusieurs variables explicatives) bien que le cadre conceptuel et les méthodes de calculs soient identiques.

Le principe de la régression linéaire est de modéliser une variable dépendante quantitative Y, au travers d'une combinaison linéaire de p variables explicatives quantitatives, $X_1, X_2, \ldots, X_p$. Le modèle déterministe (ne prenant pas en compte d'aléa) s'écrit pour une observation i,

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \epsilon_i \quad (1)$$

où y_i est la valeur observée pour la variable dépendante pour l'observation i, x_{ij} est la valeur prise par la variable j pour l'observation i, et e_i est l'erreur du modèle.

Le cadre statistique et les hypothèses qui l'accompagnent ne sont pas nécessaires pour ajuster ce modèle. Par ailleurs la minimisation par la méthode des moindres carrés (on minimise la somme des erreurs quadratiques ϵ_i^2) fournit une solution analytique exacte. Néanmoins si l'on veut pouvoir tester des hypothèses et mesurer le pouvoir explicatif des différentes variables explicatives dans le modèle, un cadre statistique est nécessaire.

Les hypothèses de la régression linéaire sont les suivantes : les erreurs ϵ_i suivent une même loi normale N(0,s) et sont indépendantes.

L'écriture du modèle complétée par cette hypothèse a pour conséquence que, dans le cadre du modèle de régression linéaire, les y_i sont des réalisations de variables aléatoires de moyenne μ_i et de variance s^2, avec

$$\mu_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} \quad (2)$$

Les estimateurs des coefficients β et de leur variance sont donnés par :

$$\hat{\beta} = (X^t X)^{-1} X^t Y \quad (3)$$

et

$$Var(\hat{\beta}) = \hat{\sigma}^2 (X^t X)^{-1} \quad (4)$$

Si l'on souhaite utiliser les différents tests proposés dans les résultats de la régression linéaire il est recommandé de vérifier a posteriori que les hypothèses sous-jacentes sont bien vérifiées. La normalité des résidus peut être vérifiée en analysant certains graphiques ou en utilisant un test de normalité. L'indépendance des résidus peut être vérifiée en analysant certains graphiques ou en utilisant le test de Durbin Watson disponible dans les outils d'analyse de séries chronologiques de XLSTAT.

Correction de l'hétéroscédasticité et de l'autocorrélation

L'homoscédasticité et l'indépendance des résidus (termes d'erreur) sont des hypothèses clefs de la régression, où, pour rappel, on suppose qu'ils sont indépendants, identiquement distribués suivant une loi normale de moyenne nulle et de variance fixe. Lorsque ces hypothèses ne peuvent être validées (un test de Durbin Watson ou de White disponibles dans les outils d'analyse des séries chronologiques permettent de les vérifier), une conséquence est que la matrice de covariance ne peut être calculée suivant la formule (4). Les variances des coefficients β peuvent alors être fausses et les conclusions quant à la significativité de la contribution ou non au modèle des variables correspondantes peuvent alors être faussées, de même que les intervalles de confiance. Une variable explicative pourrait être déclarée comme inutile alors que sa contribution est significative. XLSTAT permet de corriger les matrices de covariance pour les effets d'hétéroscédasticité et d'autocorrélation qui peuvent survenir notamment dans des cas où le temps intervient (séries chronologiques, données longitudinales).

Pour ce qui concerne l'hétéroscédasticité, White (1980) suivi par plusieurs auteurs, a exploré plusieurs façons d'améliorer le calcul de la matrice de variance-covariance des paramètres β, en prenant en compte les résidus et les leverages centrés obtenus à partir des calculs standards de la régression linéaire (voir MacKinnon (1985) et Zeileis (2006) pour une revue exhaustive). Lorsque les hypothèses de la régression linéaire ne peuvent être conservées, si les estimateurs des coefficients ne sont pas modifiés, la simplification permettant d'aboutir à l'équation (2) n'est plus possible, et l'on doit revenir à l'expression générale suivante :

$$Var(\beta) = (X^t X)^{-1} (X^t \Omega X) (X^t X)^{-1} \quad (5)$$

L'équation (5) est équivalente à l'équation (4) lorsque

$$\Omega = \hat{\sigma}^2 I \quad (6)$$

Soient ω_i les éléments de la diagonale de W. Les différents estimateurs consistents pour l'hétéroscédasticité proposés (HC, heteroscedasticity consistent) pour les ω_i sont donnés par :

$$HC0 : \omega_i = {\hat{e_i}}^2$$
$$HC1 : \omega_i = {\hat{e_i}}^2 \frac{n}{(n-p-1)}$$
$$HC2 : \omega_i = \frac{{\hat{e_i}}^2}{(1-h_i)}$$
$$HC3 : \omega_i = \frac{{\hat{e_i}}^2}{(1-h_i)^2}$$
$$HC4 : \omega_i = \frac{{\hat{e_i}}^2}{(1-h_i)^{\delta_i}} \text{ avec } \delta_i = min(4, h_i/\bar{h})$$

Où les $\hat{e_i}$ sont les résidus, et les h_i sont les *leverages* centrés, et p est le nombre de variables explicatives.

Newey et West (1987) ont suggéré un estimateur qui permet d'appliquer une correction à la fois pour l'autocorrélation et pour l'hétéroscédasticité, mais le décalage (lag) doit être connu de l'utilisateur (les outils d'analyse descriptive des séries chronologiques ou ARIMA de XLSTAT peuvent être utilisés pour cela). Pour un décalage de 0 (pas d'autocorrélation) nous avons :

$$X^t \Omega X = X^t \Omega_0 X = \frac{n}{n-p-1} \sum_{i=1}^{n} \hat{e}_i^2 x_i^t x_i$$

où x_i est le vecteur des variables explicatives (incluant un 1 pour l'intercept du modèle) pour la ième observation. Pour un décalage de m pas (m>0), nous avons :

$$X^t \Omega X = X^t \Omega_0 X + \frac{n}{n-p-1} \sum_{l=1}^{m} \sum_{t=l+1}^{n} \hat{e}_t^2 \hat{e}_{t-l}^2 (x_t^t x_{t-l} - x_{t-l}^t x_t)$$

La version inajustée de l'estimateur de Newey et West correspond à la même approche sans le facteur de correction $n/(n-p-1)$.

L'option **Classes** permet de corriger le problème d'hétéroscedasticité dans le cas où l'on considère que les variances sont égales uniquement à l'intérieur de groupes donnés. Lorsque cette option est sélectionnée, vous devez ensuite sélectionner les données indiquant à quel classe appartient chaque observation.

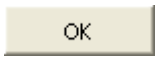
$$X^t \Omega X = \frac{n-1}{n-p-1} \frac{K}{K-1} \sum_{g=1}^{K} X_g^t \hat{\mathbf{e}} \hat{\mathbf{e}}^{\mathbf{t}} X_g$$

où K est le nombre de classes et X_g est le sous-ensemble d'observations correspondant à la classe g.

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-

dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.


: cliquez sur ce bouton pour changer la façon dont vous chargez les données dans XLSTAT. Si le bouton possède une icône de souris, XLSTAT vous permet de faire une sélection à la souris des données. Si le bouton possède une icône de liste, XLSTAT vous permet de sélectionner les données à partir d'une liste. Si le bouton possède une icône orange, des boutons complémentaires avec un point d'interrogation  sont affichés pour vous permettre d'importer les données à partir de fichiers.

Onglet **Général** :

Y / Variables dépendantes :

Quantitatives : sélectionnez la ou les variables réponse que vous souhaitez modéliser. Si plusieurs variables sont sélectionnées, XLSTAT fera les calculs pour chacune des variables indépendamment. Si des en-têtes de colonnes ont été sélectionnés, veuillez vérifier que l'option « Libellés des variables » est activée.

X / Variables explicatives :

Quantitatives : sélectionnez la ou les variables qualitatives explicatives sur la feuille Excel. Les données sélectionnées doivent être de type numérique. Si le libellé des variables a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

Qualitatives : activez cette option si vous voulez inclure une ou plusieurs variables explicatives qualitatives dans le modèle. Dans ce cas, vous ne ferez plus de la régression linéaire, mais de l'ANCOVA. Sélectionnez alors la ou les variables correspondantes sur la feuille Excel. Les données sélectionnées peuvent être de tout type, mais les données numériques sont automatiquement considérées comme nominales. Si le libellé des variables a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des variables : activez cette option si la première ligne des données sélectionnées (variables dépendantes et explicatives, libellés des observations, poids) contient un libellé.

Libellés des observations : activez cette option si vous voulez utiliser des libellés d'observations pour l'affichage des résultats. Si l'option « Libellés des variables » est activée, la première cellule de la sélection doit comprendre un en-tête. Si vous n'activez pas cette option, des libellés seront automatiquement créés (Obs1, Obs2, …).

Poids des observations : activez cette option si vous voulez pondérer les observations. Si vous n'activez pas cette option, les poids seront tous considérés comme valant 1. Les poids doivent être impérativement supérieurs ou égaux à 0. Un poids de 2 est équivalent à répéter deux fois la même observation. Si un en-tête de colonne a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

Poids dans la régression : activez cette option si vous voulez effectuer une régression par les moindres carrés pondérés. Si vous n'activez pas cette option, les poids seront tous considérés comme valant 1. Les poids doivent être impérativement supérieurs ou égaux à 0. Si un en-tête de colonne a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

Groupes : activez cette option puis sélectionnez ici les données d'appartenance à des groupes si vous souhaitez que les calculs soient effectués sur chaque groupe séparément.

Onglet **Options** :

Sous-onglet **Modèle** :

Constante fixée : activez cette option pour fixer la constante du modèle de régression à une valeur que vous devez ensuite saisir (0 par défaut).

Tolérance : activez cette option pour permettre à l'algorithme de calcul de la régression OLS ne pas prendre en compte les variables qui seraient soit constantes soit trop corrélées avec d'autres variables déjà utilisées dans le modèle (0.0001 par défaut).

Interactions / Niveau : activez cette option pour inclure des interactions dans le modèle puis entrez le niveau maximum d'interaction (valeur comprise entre 1 et 4).

Intervalle de confiance (%) : entrez l'étendue en pourcentage de l'intervalle de confiance à utiliser pour les différents tests, et pour le calcul des intervalles de confiance autour des paramètres et des prédictions. Valeur par défaut : 95.

Sélection du modèle : activez cette option si vous souhaitez utiliser l'une des quatre méthodes de sélection proposées :

- **Meilleur modèle** : cette méthode permet de choisir le meilleur modèle parmi tous les modèles comprenant un nombre de variables variant de « Min variables » à « Max variables ». Par ailleurs le « critère » pour déterminer le meilleur modèle peut être choisi par l'utilisateur.

- **Critère** : veuillez choisir le critère parmi la liste suivante : R² ajusté, Moyenne des Carrés des Erreurs (MCE), Cp de Mallows, AIC de Akaike, SBC de Schwarz, PC d'Amemiya.

- **Min variables** : entrez le nombre minimum de variables à prendre en compte dans le modèle.

- **Max variables** : entrez le nombre maximum de variables à prendre en compte dans le modèle.

Remarque : cette méthode peut entraîner des calculs longs car le nombre total de modèles explorés est la somme des (Cn, k) pour k variant entre « Min variables » et « Max variables », où (Cn, k) vaut $\frac{n!}{(n-k)!k!}$. Il est donc conseillé d'augmenter progressivement la valeur de « Max variables ».

- **Stepwise** : le processus de sélection commence par l'ajout de la variable ayant la plus forte contribution au modèle (le critère utilisé est la statistique$t$de Student). Si une seconde variable est telle que la probabilité associée à son t est inférieure à la « **Probabilité pour l'entrée** », elle est ajoutée au modèle. De même pour une troisième variable. A partir de l'ajout de la troisième variable, après chaque ajout, on évalue pour toutes les variables présentes dans le modèle quel serait l'impact de son retrait (toujours au travers de la statistique t). Si la probabilité est supérieure à la « **Probabilité pour le retrait** », la variable est retirée. La procédure se poursuit jusqu'à ce que plus aucune variable ne puisse être ajoutée/retirée.

- **Ascendante** : la procédure est identique à celle de la sélection progressive, hormis le fait que les variables sont uniquement ajoutées et jamais retirées.

- **Descendante** : la procédure commence par l'ajout simultané de toutes les variables. Les variables sont ensuite retirées du modèle suivant la procédure utilisée pour la sélection progressive.

Sous-onglet **Covariances** :

Dans cet onglet, vous pouvez choisir d'appliquer des corrections pour l'hétéroscedasticité et l'autocorrélation. Veuillez vous reporter à la section *Description* pour plus de détails.

Onglet **Validation** :

Validation : activez cette option si vous souhaitez utiliser une partie des données sélectionnées pour valider le modèle.

Jeu de validation : choisissez l'une des options pour définir le mode de sélection des observations utilisées pour la validation :

- **Aléatoire** : les observations sont sélectionnées de manière aléatoire. Le « Nombre d'observations » doit alors être saisi.
- **N dernières lignes** : les N dernières observations sont sélectionnées pour la validation. Le « Nombre d'observations » N doit alors être saisi.
- **N premières lignes** : les N premières observations sont sélectionnées pour la validation. Le « Nombre d'observations » N doit alors être saisi.
- **Variable de groupe** : si vous choisissez cette option, vous devez ensuite sélectionner une variable indicatrice composée de 0 pour les observations à utiliser pour le calcul du modèle, et de 1 pour les observations à utiliser pour la validation du modèle.

Onglet **Prédiction** :

Prédiction : activez cette option si vous souhaitez sélectionner des données à utiliser en mode prédiction. Si vous activez cette option, vous devez veiller à ce que les données de prédiction soient organisées comme les données d'estimation : mêmes variables, même ordre dans les sélections. En revanche vous ne devez pas sélectionner de libellés de variables : la première ligne des sélections décrites ci-dessous doit être une ligne de données.

Quantitatives : activez cette option pour sélectionner la ou les variables quantitatives explicatives. La première ligne ne doit pas comprendre d'en- tête.

Qualitatives : activez cette option pour sélectionner la ou les variables qualitatives explicatives. La première ligne ne doit pas comprendre d'en-tête.

Libellés des observations : activez cette option si vous voulez utiliser des libellés d'observations disponibles sur une feuille Excel pour l'affichage des résultats. La première ligne ne doit pas comprendre d'en-tête. Si vous n'activez pas cette option, des libellés seront automatiquement créés (PredObs1, PredObs2, …).

Onglet **Données manquantes** :

Supprimer les observations : activez cette option pour supprimer les observations comportant des données manquantes.

- **Vérifier pour chaque Y séparément** : choisissez cette option si vous voulez que lorsque, pour une observation donnée, il y a des données manquantes uniquement dans les Y, l'observation ne soit supprimée que si la donnée correspondant au Y en cours de modélisation est manquante.
- **Pour tous les Y** : choisissez cette option pour supprimer toutes les observations pour lesquelles des Y sont manquants.
- Remarque : les deux alternatives ci-dessus sont sans effet s'il n'y a qu'un seul Y.

Ignorer les données manquantes : si vous choisissez cette option, pour les données manquantes correspondant aux variables dépendantes XLSTAT essaiera de les estimer à partir du modèle obtenu. Pour celles correspondant aux variables explicatives, les observations

correspondantes seront conservées dans la mesure du possible pour estimer la matrice de variance covariance (suppression par paire).

Estimer les données manquantes : activez cette option pour estimer les données manquantes avant le début des calculs.

- **Moyenne ou mode** : activez cette option pour estimer les données manquantes en utilisant la moyenne (variables quantitatives) ou le mode (variables qualitatives) pour les variables correspondantes.
- **Plus proche voisin** : activez cette option pour estimer les données manquantes d'une observation en recherchant le plus proche voisin de l'observation.

Onglet **Sorties** :

Sous-onglet **Général** :

Statistiques descriptives : activez cette option pour afficher les statistiques descriptives pour les variables sélectionnées.

Corrélations : activez cette option pour afficher la matrice de corrélation pour les variables quantitatives (dépendantes et explicatives).

Statistiques de multicolinéarité : activez cette option pour afficher les statistiques de multicolinéarité.

Mesures de la taille de l'effet : activez cette option pour afficher les mesures de la taille de l'effet. Dans le cadre de la régression linéaire, les mesures suivantes sont affichées : * R^2 * $f^2 = \frac{R^2}{1-R^2}$

Analyse de la variance : activez cette option pour afficher le tableau d'analyse de la variance.

Type I/III SS : activez cette option pour afficher les tableaux Type I SS et Type III SS permettant de mesurer la contribution des différentes variables explicatives au modèle (SS correspond à *Sum of Squares*).

Press : activez cette option pour calculer et afficher la statistique Press (predicted residual error sum of squares).

Interprétation : activez cette option pour que XLSTAT calcule une interprétation automatique des résultats.

Coefficients normalisés : activez cette option pour afficher les paramètres normalisés du modèle (coefficients bêta).

Prédictions et résidus : activez cette option pour afficher les prédictions et les résidus pour l'ensemble des observations.

- **X** : activez cette option pour afficher dans le tableau des prédictions, pour chaque observation, les données correspondant aux différentes variables explicatives

quantitatives.

- **Intervalles de confiance** : activez cette option pour calculer et afficher les intervalles de confiance sur les prédictions.
- **Prédictions ajustées** : activez cette option pour calculer et afficher les prédictions ajustées dans le tableau des prédictions et résidus.
- **Diagnostics d'influence** : activez cette option pour calculer et afficher le tableau des statistiques permettant d'identifier les observations ayant une influence sur les prédictions ou sur les coefficients associés à certaines variables explicatives (voir section résultats).

Sous-onglet **Contrastes** :

Calculer contrastes : activer cette option pour calculer les contrastes, puis sélectionnez le tableau des contrastes, où il doit y avoir une colonne par contraste et une ligne pour chaque coefficient du modèle.

Sous-onglet **Test des hypothèses** :

Cette option n'est disponible que si dans l'onglet Sorties/Général, l'option **prédictions et résidus** est activée.

Test de normalité : activez cette option pour qu'un test de Shapiro-Wilk soit effectué sur les résidus.

Onglet **Graphiques** :

Graphiques de régression : activez cette option pour afficher les graphiques de régression :

- **Coefficients normalisés** : activez cette option pour afficher sur un graphique les paramètres normalisés du modèle avec leur intervalle de confiance.
- **Prédictions et résidus** : activez cette option pour afficher les graphiques suivants :

(1) Droite de régression : ce graphique n'est affiché que s'il n'y a qu'une seule variable explicative, et que cette variable est quantitative.

(2) Variable explicative versus résidus normalisés : ce graphique n'est affiché que s'il n'y a qu'une seule variable explicative, et que cette variable est quantitative.

(3) Variable dépendante versus résidus normalisés.

(4) Prédictions pour la variable dépendante versus variable dépendante.

(5) Graphique en bâtons des résidus normalisés.

- **Intervalles de confiance** : activez cette option pour afficher les intervalles de confiance sur les graphiques (1) et (4).

Résultats

Statistiques descriptives : les tableaux de statistiques descriptives présentent pour toutes les variables sélectionnées des statistiques simples. Pour les variables dépendantes et les variables explicatives quantitatives, sont affichés le nombre d'observations, le nombre de données manquantes, le nombre de données non manquantes, la moyenne, et l'écart-type (non biaisé).

Matrice de corrélation : ce tableau est affiché afin de vous permettre d'avoir un aperçu des corrélations entre les différentes variables sélectionnées.

Synthèse de la sélection des variables : dans le cas où une méthode de sélection a été choisie, XLSTAT affiche la synthèse de la sélection. Dans le cas d'une sélection pas à pas, les statistiques correspondant aux différentes étapes sont affichées. Dans le cas d'une sélection du meilleur modèle pour un nombre de variables variant de p à q, le meilleur modèle pour chaque nombre de variable est affiché avec les statistiques correspondantes ; le meilleur modèle pour le critère choisi est alors affiché en gras.

Coefficients d'ajustement : dans ce tableau sont affichées les statistiques relatives à l'ajustement du modèle de régression :

- **Observations** : le nombre d'observations prises en compte dans les calculs. Dans les formules présentées ci-dessous n désigne le nombre d'observations.
- **Somme des poids** : la somme des poids des observations prises en compte dans les calculs. Dans les formules présentées ci-dessous W désigne la somme des poids.
- **DDL** : le nombre de degrés de liberté pour le modèle retenu (correspondant à la partie erreurs).
- **R²** : le coefficient de détermination du modèle. La valeur de ce coefficient est comprise entre 0 et 1. XLSTAT le calcule comme suit :

$$R^2 = \frac{\sum_{i=1}^{n} w_i(\hat{y}_i - \bar{y}_i)^2}{\sum_{i=1}^{n} w_i(y_i - \bar{y}_i)^2} \text{ avec } \bar{y} = \frac{1}{W}\sum_{i=1}^{n} w_i y_i$$

Le R^2 s'interprète comme la proportion de la variabilité de la variable dépendante expliquée par le modèle. Plus le R^2 est proche de 1, meilleur est le modèle. L'inconvénient du R^2 est qu'il ne prend pas en compte le nombre de variables utilisées pour ajuster le modèle.

- **R² ajusté** : le coefficient de détermination ajusté du modèle. Le R^2 ajusté peut être négatif si le R^2 est voisin de zéro. Ce coefficient n'est affiché que si la constante du modèle n'est pas fixée par l'utilisateur. Sa valeur est définie par

$$\hat{R}^2 = 1 - (1 - R^2)\frac{W - 1}{W - p - 1}$$

Le R^2 ajusté est une correction du R^2 qui permet de prendre en compte le nombre de variables utilisées dans le modèle.

- **MCE** : la moyenne des carrés des erreurs (MCE) est définie par :

$$MCE = \frac{1}{W - p^*} \sum_{i=1}^{n} w_i (y_i - \hat{y}_i)^2$$

- **RMCE** : la racine de la moyenne des carrés des erreurs (RMCE) est la racine carrée de la MCE.

- **MAPE** : la *Mean Absolute Percentage Error* est calculée comme suit :

$$MAPE = \frac{100}{W} \sum_{i=1}^{n} w_i \left| \frac{y_i - \hat{y}_i}{y_i} \right|$$

- **DW** : le coefficient de Durbin-Watson est défini par

$$DW = \frac{\sum_{i=2}^{n} [(y_i - \hat{y}_i) - (y_{i-1} - \hat{y}_{i-1})]^2}{\sum_{i=1}^{n} w_i (y_i - \hat{y}_i)^2}$$

Ce coefficient correspond au coefficient d'autocorrélation d'ordre 1 et permet de vérifier que les résidus du modèle ne sont pas autocorrélés, sachant que l'indépendance des résidus est l'une des hypothèses de base de la régression linéaire. L'utilisateur pourra se référer à une table des coefficients de Durbin-Watson pour vérifier si l'hypothèse d'indépendance des résidus est acceptable.

- **Cp** : le coefficient Cp de Mallows est défini par

$$Cp = \frac{SCE}{\hat{\sigma}} + 2p^* - W$$

où SCE est la somme du carré des erreurs pour le modèle avec p variables explicatives, et où $\hat{\sigma}$ correspond à l'estimateur de la variance des résidus pour le modèle comprenant toutes les variables explicatives. Plus le coefficient Cp est proche de p* moins le modèle est biaisé.

- **AIC** : le critère d'information d'Akaike (Akaike's Information Criterion) est défini par

$$AIC = W \ln(\frac{SCE}{W}) + 2p^*$$

Ce critère proposé par Akaike (1973) dérive de la théorie de l'information, et s'appuie sur la mesure de Kullback et Leibler (1951). C'est un critère de sélection de modèles qui pénalise les modèles pour lesquels l'ajout de nouvelles variables explicatives n'apporte pas suffisamment d'information au modèle, l'information étant mesurée au travers de la SCE. On cherche à minimiser le critère AIC.

- **SBC** : le critère bayésien de Schwarz (Schwarz's Bayesian Criterion) est défini par

$$SBC = W \ln(\frac{SCE}{W}) + \ln(W)p^*$$

Ce critère proposé par Schwarz (1978) est proche du critère AIC, et comme ce dernier on cherche à le minimiser.

- **PC** : le critère de prédiction d'Amemiya (Amemiya's Prediction Criterion) est défini par

$$PC = \frac{(1 - R^2)(W + p^*)}{W - p^*}$$

Ce critère proposé par Amemiya (1980) permet comme le R² ajusté de tenir compte de la parcimonie du modèle.

- **Press** : la statistique du Press (predicted residual error sum of squares) n'est affichée que si l'option correspondante a été activée dans la boîte de dialogue. Elle est définie par

$$Press = \sum_{i=1}^{n} w_i (y_i - \hat{y}_{i(-i)})^2$$

où $\hat{y}_{i(-i)}$ est la prédiction pour l'observation i lorsque cette dernière n'est pas utilisée pour l'estimation des paramètres. On obtient alors

$$\text{Press RMCE } = \sqrt{\frac{\text{Press}}{W - p^*}}$$

Le Press RMCE peut alors être comparé au RMCE. Une différence importante entre les deux indique que le modèle est sensible à la présence ou absence de certaines observations dans le modèle.

- **Q²** : cette statistique aussi connue comme le R² de validation croisée, n'est affichée que si l'option Press est activée. Elle est définie par :

$$Q^2 = 1 - \frac{\text{Press}}{\sum_{i=1}^{n} (y_i - \overline{y})^2}$$

Le Q² indique la proportion de la variance totale expliquée par les variables explicatives lorsque les prédictions pour chaque observation sont calculées lorsque l'observation en question n'est pas dans l'échantillon servant à l'estimation des paramètres. Une différence importante entre le Q² et le R² indique que le modèle est sensible à l'ajout ou au retrait de certaines observations dans l'échantillon d'estimation.

Le **tableau d'analyse de la variance** permet d'évaluer le pouvoir explicatif des variables explicatives. Dans le cas où la constante du modèle n'est pas fixée à une valeur donnée, le pouvoir explicatif est évalué en comparant l'ajustement (au sens des moindres carrés) du modèle final avec l'ajustement du modèle rudimentaire composé d'une constante égale à la moyenne de la variable dépendante. Dans le cas où la constante du modèle est fixée, la comparaison est faite par rapport au modèle pour lequel la variable dépendante serait égale à la constante fixée.

Si l'option Type I/III SS (SS : Sum of Squares) est activée, les tableaux suivants sont affichés.

Le tableau des **Type I SS** (sommes de carrés de type I) permet de visualiser l'influence de l'ajout progressif des variables explicatives sur l'ajustement du modèle, au sens de la somme des carrés des erreurs (SCE), de la moyenne des carrés des erreurs (MCE), du F de Fisher, ou de la probabilité associée au F de Fisher. Plus la probabilité est faible, plus la contribution de la variable au modèle est importante, toutes les autres variables étant déjà dans le modèle.

Remarques : l'ordre de sélection des variables dans le modèle influe sur les valeurs obtenues ; la somme des sommes des carrés de ce tableau est égale à la somme des carrés du modèle.

Le tableau des **Type II SS** (sommes de carrés de type II) permet de visualiser l'influence du retrait d'une variable explicative et des interactions qui en dépendent, sur l'ajustement du modèle, toutes les autres variables étant conservées, au sens de la somme des carrés des erreurs (SCE), de la moyenne des carrés des erreurs (MCE), du F de Fisher, ou de la probabilité associée au F de Fisher. Plus la probabilité est faible, plus la contribution de la variable au modèle est importante, toutes les autres variables étant déjà dans le modèle. Remarque : contrairement au cas des Type I SS, l'ordre de sélection des variables dans le modèle n'influe pas sur les valeurs obtenues.

Le tableau des **Type III SS** (sommes de carrés de type III) permet de visualiser l'influence du retrait d'une variable explicative sur l'ajustement du modèle, toutes les autres variables étant conservées (y compris les interactions incluant la variable en question), au sens de la somme des carrés des erreurs (SCE), de la moyenne des carrés des erreurs (MCE), du F de Fisher, ou de la probabilité associée au F de Fisher. Plus la probabilité est faible, plus la contribution de la variable au modèle est importante, toutes les autres variables étant déjà dans le modèle. Remarque : contrairement au cas des Type I SS, l'ordre de sélection des variables dans le modèle n'influe pas sur les valeurs obtenues.

Le tableau des **paramètres du modèle** affiche l'estimation des paramètres, l'écart-type correspondant, le t de Student, la probabilité correspondante, ainsi que l'intervalle de confiance. Si l'intervalle de confiance comprend 0, alors le poids d'une variable dans le modèle n'est pas significatif.

L'**équation du modèle** est ensuite affichée pour faciliter la lecture ou la réutilisation du modèle.

Le tableau des **coefficients normalisés** (aussi appelés coefficients bêta) permet de comparer le poids relatif des variables. Plus la valeur absolue d'un coefficient est élevée, plus le poids de la variable correspondante est important. Lorsque l'intervalle de confiance autour des coefficients normalisés comprend la valeur 0 (cela est facilement visible sur le graphique des coefficients normalisés), le poids d'une variable dans le modèle n'est pas significatif.

Dans le tableau des **prédictions et résidus** sont donnés pour chaque observation, son poids, la valeur observée de la variable dépendante, la prédiction du modèle, les résidus, les résidus studentisés, les intervalles de confiance, ainsi que la prédiction ajustée si l'option correspondante a été activées dans la boîte de dialogue. Deux types d'intervalles de confiance sont affichés : un intervalle de confiance autour de la moyenne (correspondant au cas où l'on ferait la prédiction pour un nombre infini d'observations avec un ensemble de valeurs données des variables explicatives) et un intervalle autour de la prédiction ponctuelle (correspondant au cas d'une prédiction isolée pour des valeurs données des variables explicatives). Le second intervalle est toujours plus grand que le premier, l'incertitude étant plus importante. Si des données de validation ont été sélectionnées, elles sont affichées en fin de tableau.

Les **graphiques** qui suivent permettent de visualiser les résultats mentionnés ci-dessus. S'il n'y a qu'une seule variable explicative dans le modèle, le premier graphique affiché permet de visualiser les valeurs observées, la droite de régression et les deux types d'intervalles de confiance autour des prévisions. Le second graphique permet quant à lui de visualiser les résidus normalisés en fonction de la variable explicative. En principe, les résidus doivent être distribués de manière aléatoire autour de l'axe des abscisses. L'observation d'une tendance ou d'une forme révèlerait un problème au niveau du modèle.

Les **trois graphiques** affichés ensuite permettent de visualiser respectivement l'évolution des résidus normalisés en fonction de la variable dépendante, la distance entre les prédictions et les observations (pour un modèle idéal, les points seraient tous sur la bissectrice), et les résidus normalisés sur la forme d'un diagramme en bâtons. Ce dernier graphique permet de rapidement voir si un nombre anormal de données sort de l'intervalle]-2, 2[sachant que ce dernier, sous hypothèse de normalité, doit contenir environ 95% des données.

Dans le tableau des **diagnostics d'influence** sont affichés pour chaque observation, son poids, le résidu, le résidu normalisé (division par la RMCE), le résidu studentisé, le résidu supprimé (Deleted), le résidu supprimé studentisé, le leverage centré, la distance de Mahalanobis, le D de Cook, le CovRatio, le DFFits, le DFFits standardisé, les DFBeta (un par coefficient du modèle) et les DFBeta standardisés.

Quatre graphiques sont ensuite affichés pour mettre en évidence les observations dont l'influence nécessite une analyse particulière.

Si vous avez sélectionné des données à utiliser pour calculer des **prédictions sur de nouvelles observations**, le tableau correspondant est ensuite affiché.

Exemple

Un exemple de régression linéaire simple est disponible sur le Centre d'aide XLSTAT à l'adresse

http://www.xlstat.com/demo-regf.htm

Un exemple de régression linéaire multiple est disponible sur le Centre d'aide XLSTAT à l'adresse

http://www.xlstat.com/demo-reg2f.htm

Bibliographie

Akaike H. (1973). Information theory and the extension of the maximum likelihood principle. In: Second International Symposium on Information Theory. (Eds: V.N. Petrov and F. Csaki). Academiai Kiadó, Budapest. 267-281.

Amemiya T. (1980). Selection of regressors. *International Economic Review*, **21**, 331-354.

Cook R.D. (1977). Detection of influential observations in linear regression. *Technometrics*, **19**, 15-18.

Dempster A.P. (1969). Elements of Continuous Multivariate Analysis. Addison-Wesley, Reading.

Jobson J. D. (1999). Applied Multivariate Data Analysis: Volume 1: Regression and Experimental Design. Springer Verlag, New York.

Mallows C.L. (1973). Some comments on Cp. *Technometrics*, **15**, 661-675.

Rogers W. H. (1993). Regression standard errors in clustered samples. _Stata Technical _Bulletin, **13**, 19–23.

Tomassone R., Audrain S., Lesquoy de Turckheim E. and Miller C. (1992). La Régression, Nouveaux Regards sur une Ancienne Méthode Statistique. INRA et MASSON, Paris.

Velleman P.F. and R.E. Welsch (1981). Efficient computing of regression diagnostics. *The American Statistician*, **35**, 234-242.

Welsch R.E. and Kuh E. (1977). Linear Regression Diagnostics. *Sloan School of Management Working Paper*,923-977, M.I.T., Cambridge, Mass.

White H. (1980). A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity. *Econometrica*, **48(4)**, 817-838.

Zeileis A. (2006). Object-Oriented Computation of Sandwich Estimators. *Journal of Statistical Software*,**16(9)**, 1-16.

ANOVA

Utilisez ce module pour faire de l'ANOVA (Analyse de variance) à un ou plusieurs facteurs, équilibrée ou déséquilibrée. Des options avancées vous permettent de choisir les contraintes sur le modèle et de tenir compte des interactions entre les facteurs. Des tests de comparaisons multiples peuvent être calculés.

Dans cette section :

Description

L'analyse de variance utilise le même cadre conceptuel que la régression linéaire. La différence principale vient de la nature des variables explicatives : au lieu d'être quantitatives, elles sont ici qualitatives. Dans le cadre de l'ANOVA, les variables explicatives sont souvent appelées facteurs.

Si p est le nombre de facteurs, le modèle de l'ANOVA s'écrit de la manière suivante :

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_{k(i,j),j} + \epsilon_i \tag{1}$$

où y_i est la valeur observée pour la variable dépendante pour l'observation i, k(i,j) est l'indice correspondant à la modalité du facteur j pour l'observation i, et ϵ_i est l'erreur du modèle.

Les hypothèses utilisées en ANOVA sont identiques à celles de la régression linéaire : les erreurs ϵ_i suivent une même loi normale $N(0, s)$ et sont indépendantes.

L'écriture du modèle complétée par cette hypothèse a pour conséquence que, dans le cadre du modèle de régression linéaire, les y_i sont des réalisations de variables aléatoires de moyenne μ_i et de variance s^2, avec

$$\mu_i = \beta_0 + \sum_{j=1}^{p} \beta_{k(i,j),j} \tag{2}$$

Les estimateurs des coefficients β et de leur variance sont donnés par :

$$\hat{\beta} = (X^t X)^{-1} X^t Y \tag{3}$$

et

$$Var(\hat{\beta}) = \hat{\sigma}^2 (X^t X)^{-1} \quad (4)$$

Si l'on souhaite utiliser les différents tests proposés dans les résultats de la régression linéaire il est recommandé de vérifier a posteriori que les hypothèses sous-jacentes sont bien vérifiées. La normalité des résidus peut être vérifiée en analysant certains graphiques ou en utilisant un test de normalité. L'indépendance des résidus peut être vérifiée en analysant certains graphiques ou en utilisant le test de Durbin Watson disponible dans les outils d'analyse de séries chronologiques de XLSTAT. L'homoscédasticité peut être vérifiée au travers d'un test de Levene.

Sélection des données

Dans XLSTAT, il est possible de rentrer les données de deux façons différentes lorsque l'on a une variable dépendante et jusqu'à maximum trois facteurs :

- Chaque variable est entrée sous forme de colonne.
- On rentre un tableau où les lignes définissent les données en fonction d'un facteur, et les colonnes les définissent en fonction des autres facteurs (1 ou 2 variables). Dans ce cas, il est impossible d'estimer les données manquantes. Elles seront donc automatiquement supprimées.

Correction de l'hétéroscédasticité et de l'autocorrélation

L'homoscedasticité et l'indépendance des résidus (termes d'erreur) sont des hypothèses clefs de la régression, où, pour rappel, on suppose qu'ils sont indépendants, identiquement distribués suivant une loi normale de moyenne nulle. Lorsque ces hypothèses ne peuvent être validées (un test de Durbin Watson ou de White disponibles dans les outils d'analyse des séries chronologiques permettent de les vérifier), une conséquence est que la matrice de covariance ne peut être calculée suivant la formule (4). Les variances des coefficients *b* peuvent alors être fausses et les conclusions quant à la significativité de la contribution ou non au modèle des variables correspondantes peuvent alors être faussées, de même que les intervalles de confiance. Une variable explicative pourrait être déclarée comme inutile alors que sa contribution est significative. XLSTAT permet de corriger les matrices de covariance pour les effets d'hétéroscédasticité et d'autocorrélation qui peuvent survenir notamment dans des cas où le temps intervient (séries chronologiques, données longitudinales).

Pour ce qui concerne l'hétéroscédasticité, White (1980) suivi par plusieurs auteurs, a exploré plusieurs façons d'améliorer le calcul de la matrice de variance-covariance des paramètres β , en prenant en compte les résidus et les leverages centrés obtenus à partir des calculs standards de la régression linéaire (voir MacKinnon (1985) et Zeileis (2006) pour une revue exhaustive). Lorsque les hypothèses de la régression linéaire ne peuvent être conservées, si les estimateurs des coefficients ne sont pas modifiés, la simplification permettant d'aboutir à l'équation (4) n'est plus possible, et l'on doit revenir à l'expression générale suivante :

$$Var(\beta) = (X^t X)^{-1} (X^t \Omega X) (X^t X)^{-1} \quad (5)$$

L'équation (5) est équivalente à l'équation (4) lorsque

$$\Omega = \hat{\sigma}^2 I \quad (6)$$

Soient ω_i les éléments de la diagonale de W. Les différents estimateurs consistents pour l'hétéroscédasticité proposés (HC, heteroscedasticity consistent) pour les ω_i sont donnés par :

$$
\begin{aligned}
HC0: &\quad \omega_i = {\hat{e_i}}^2 \\
HC1: &\quad \omega_i = {\hat{e_i}}^2 \frac{n}{(n-p-1)} \\
HC2: &\quad \omega_i = \frac{{\hat{e_i}}^2}{(1-h_i)} \\
HC3: &\quad \omega_i = \frac{{\hat{e_i}}^2}{(1-h_i)^2} \\
HC4: &\quad \omega_i = \frac{{\hat{e_i}}^2}{(1-h_i)^{\delta_i}} \text{ avec } \delta_i = min(4, h_i/\bar{h})
\end{aligned}
$$

Où les $\hat{e_i}$ sont les résidus, et les h_i sont les *leverages* centrés, et p est le nombre de variables explicatives.

Newey et West (1987) ont suggéré un estimateur qui permet d'appliquer une correction à la fois pour l'autocorrélation et pour l'hétéroscédasticité, mais le décalage (lag) doit être connu de l'utilisateur (les outils d'analyse descriptive des séries chronologiques ou ARIMA de XLSTAT peuvent être utilisés pour cela). Pour un décalage de 0 (pas d'autocorrélation) nous avons :

$$X^t \Omega X = X^t \Omega_0 X = \frac{n}{n-p-1} \sum_{i=1}^{n} \hat{e}_i^2 x_i^t x_i$$

où x_i est le vecteur des variables explicatives (incluant un 1 pour l'intercept du modèle) pour la ième observation. Pour un décalage de m pas (m>0), nous avons :

$$X^t \Omega X = X^t \Omega_0 X + \frac{n}{n-p-1} \sum_{l=1}^{m} \sum_{t=l+1}^{n} \hat{e}_t^2 \hat{e}_{t-l}^2 (x_t^t x_{t-l} - x_{t-l}^t x_t)$$

La version inajustée de l'estimateur de Newey et West correspond à la même approche sans le facteur de correction $n/(n-p-1)$.

L'option **Classes** permet de corriger le problème d'hétéroscedasticité dans le cas où l'on considère que les variances sont égales uniquement à l'intérieur de groupes donnés. Lorsque cette option est sélectionnée, vous devez ensuite sélectionner les données indiquant à quelle classe appartient chaque observation.

$$X^t \Omega X = \frac{n-1}{n-p-1} \frac{K}{K-1} \sum_{g=1}^{K} X_g^t \hat{\mathbf{e}} \hat{\mathbf{e}}^{\mathbf{t}} X_g$$

où K est le nombre de classes et X_g est le sous-ensemble d'observations correspondant à la classe g.

Interactions

On désigne par interaction un facteur artificiel (non mesuré) reflétant l'interaction entre au moins deux facteurs mesurés. Par exemple, si on applique un traitement à une plante, et que les essais sont réalisés sous deux intensités lumineuses différentes, on pourra inclure dans le modèle un facteur d'interaction traitement*lumière qui permettra d'identifier une éventuelle interaction entre les deux facteurs. S'il y a une interaction entre les deux facteurs, on observera sur les plantes un effet significativement plus important lorsque la lumière est forte et que le traitement est de type 2, alors que l'effet est moyen pour les couples (lumière faible, traitement 2) et (lumière forte, traitement 1).

Pour faire un parallèle avec la régression linéaire, les interactions sont équivalentes à des produits entre les valeurs explicatives continues, bien qu'ici l'obtention des interactions nécessite plus qu'une simple multiplication entre deux variables. Néanmoins la notation utilisée pour représenter l'interaction entre le facteur A et le facteur B est A*B.

XLSTAT permet de facilement définir les interactions à prendre en compte dans le modèle.

Facteurs imbriqués

Lorsqu'on ne peut pas croiser toutes les modalités de deux facteurs, alors on peut utiliser des facteurs imbriqués ou hiérarchiques. Par exemple, si l'on cherche à analyser le lien entre certaines caractéristiques d'un produit en sortie d'une chaîne de fabrication et les opérateurs et les machines impliqués, et si les opérateurs travaillent sur une machine donnée avec par exemple quatre opérateurs en rotation, alors chaque opérateur n'est pas croisé avec chaque machine, mais est associé à une seule machine. On a alors un effet imbriqué (l'effet opérateur est imbriqué dans l'effet machine).

XLSTAT permet d'identifier automatiquement les facteurs imbriqués. Par ailleurs, il est possible d'inclure dans le modèle un facteur imbriqué.

ANOVA équilibrée et déséquilibrée

On parle d'ANOVA équilibrée lorsque les effectifs des modalités sont égaux pour l'ensemble des facteurs. Lorsque les effectifs de toutes les modalités de l'un des facteurs ne sont pas égaux, alors l'ANOVA est dite déséquilibrée. XLSTAT permet de traiter les deux cas.

Facteurs aléatoires

Il est possible de supposer que certains facteurs sont aléatoires dans une analyse de la variance. Lorsque certains facteurs sont déclarés comme aléatoires, un nouveau tableau des carrés moyens est affiché.

Anova restreinte

Les hypothèses du modèle restreint surviennent lorsque nous avons une interaction entre un facteur fixe et un facteur aléatoire et que nous supposons que la somme des coefficients aléatoires dans le terme d'interaction au travers de chaque indice du facteur fixe vaut zéro. En résumé, la somme des effets d'interaction sur les niveaux du facteur fixe est égale à zéro.

Contraintes

Au cours des calculs, chaque facteur est décomposé en une sous-matrice contenant autant de colonnes qu'il y a de modalités dans le facteur. Typiquement, il s'agit d'un tableau disjonctif complet. Cette décomposition pose néanmoins un problème : s'il y a g modalités, le rang de cette sous- matrice n'est pas g mais g-1. Cela entraîne la nécessité de supprimer l'une des colonnes de la sous-matrice, et éventuellement de transformer les autres colonnes. Plusieurs stratégies sont possibles en fonction de l'interprétation que l'on veut ensuite faire :

1) **a1=0** : le paramètre correspondant à la première modalité est nul. Ce choix permet d'imposer que l'effet de la première modalité correspond à un standard. Dans ce cas, la constante du modèle est égale à la moyenne de la variable dépendante pour le groupe 1.

2) **an=0** : le paramètre correspondant à la dernière modalité est nul. Ce choix permet d'imposer que l'effet de la dernière modalité correspond à un standard. Dans ce cas, la constante du modèle est égale à la moyenne de la variable dépendante pour le groupe g.

3) **Somme(ai)=0** : la somme des paramètres est nulle. Ce choix permet d'imposer que la constante du modèle est égale à la moyenne de la variable dépendante lorsque l'ANOVA est équilibrée.

4) **Somme(ni.ai)=0** : la somme des paramètres est nulle. Ce choix permet d'imposer que la constante du modèle est égale à la moyenne de la variable dépendante même lorsque l'ANOVA est déséquilibrée.

Remarque : si le choix de la contrainte influence la valeur des paramètres, il n'en a aucun sur les valeurs prédites et sur les différentes statistiques d'ajustement.

Tests de comparaisons multiples

L'une des applications principales de l'ANOVA sont les tests de comparaisons multiples dont le but est de vérifier si les paramètres correspondant aux différentes modalités d'un facteur sont significativement différents ou non. Par exemple, dans le cas où quatre traitements sont appliqués à des plantes, on veut savoir non seulement si les traitements ont un effet significatif, mais aussi si les traitements ont un effet différent.

De nombreux tests ont été proposés pour comparer les moyennes des modalités. La majorité de ces tests s'appuie aussi sur l'hypothèse de normalité. XLSTAT propose les principaux tests parmi lesquels :

Test de Tukey (HSD) : ce test est le plus utilisé (HSD : *honestly significant difference*).

Test de Fisher (LSD) : c'est un test de Student qui permet de tester l'hypothèse nulle que toute les moyennes pour les différentes modalités sont égales (LSD : *least significant difference*).

Test du t* de Bonferroni : dérivé du test de Student, il est un peu plus performant car il prend en compte le fait que plusieurs comparaisons sont effectuées simultanément. En conséquence le niveau de signification du test est modifié suivant la formule suivante :

$$\alpha' = \frac{\alpha}{g(g-1)/2}$$

où g est le nombre de modalités du facteur dont les modalités sont comparées.

Test de Dunn-Sidak : dérivé du test de Bonferroni, il est plus performant dans certaines situations.

$$\alpha' = 1 - (1 - \alpha)^{1/[c(c-1)]}$$

Les tests suivants sont plus complexes et consistent en des procédures itératives pour lesquelles les résultats dépendent du nombre de combinaisons restant à tester.

Test de Newman-Keuls (SNK) : dérivé du test de Student (SNK : Student Newman-Keuls), il est très souvent utilisé bien que pas très performant.

Test de Duncan : ce test est peu utilisé.

Test de REGWQ : ce test est la procédure itérative la plus performante dans une majorité de situations (REGWQ : Ryan-Einot-Gabriel-Welsch).

La procédure de **Benjamini-Hochberg** permet de contrôler le taux de faux positifs (False Discovery Rate ou FDR). Cette procédure de pénalisation des p-values est peu conservative.

Le test de **Games-Howell** (GH) peut être utilisé dans les ANOVAs à un facteur lorsque les variances ne sont pas d'homogènes. Il peut être utilisé avec des tailles d'échantillon inégales, mais il est recommandé de l'utiliser quand le plus petit échantillon a 5 éléments ou plus, sinon il est trop libéral (au sens qu'il a tendance à rejeter trop facilement l'hypothèse H0). Le test de **T2 du Tamhane** est plus conservateur, mais pas aussi puissant que le test GH.

Tous les tests ci-dessus permettent de comparer toutes les paires de modalités et appartiennent à la famille des tests MCA (*Multiple Comparisons of All*, ou *All-Pairwise Comparisons*).

D'autres tests permettent de comparer toutes les catégories à une modalité témoin. Ces tests sont appelés tests MCB (*Multiple Comparisons with the Best*, *Comparisons with a control*). XLSTAT propose le test de Dunnett qui est le plus utilisé. On distingue trois tests de Dunnett :

- **Test bilatéral** : l'hypothèse nulle suppose l'égalité entre la modalité testée et la modalité témoin. L'hypothèse alternative suppose que les moyennes des deux modalités sont différentes.
- **Test unilatéral à gauche** : l'hypothèse nulle suppose l'égalité entre la modalité testée et la modalité témoin. L'hypothèse alternative suppose que la moyenne de la modalité témoin est inférieure à la moyenne de la modalité testée.
- **Test unilatéral à droite** : l'hypothèse nulle suppose l'égalité entre la modalité testée et la modalité témoin. L'hypothèse alternative suppose que la moyenne de la modalité témoin est supérieure à la moyenne de la modalité testée.

Tests robustes de comparaison de moyennes pour une ANOVA à un facteur

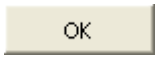
Il peut arriver que les variances ne puissent pas être supposées égales. Dans ce cas le F de l'analyse de la variance n'est pas assez robuste pour être utilisé. XLSTAT propose deux tests basés sur la distribution F mais plus robustes à l'inégalité entre les moyennes dans le cas d'une analyse de la variance à un facteur.

Ces tests sont :

- **le test de Welch ou ANOVA de Welch** (Welch,1951). Le test de Welch ajuste le dénominateur du rapport F de sorte qu'il ait la même espérance que le numérateur lorsque l'hypothèse nulle est vraie, malgré l'hétérogénéité des variances.
- **le test de Brown-Forsythe ou Rapport des F de Brown-Forsythe** (1974). Ce test utilise un dénominateur différent pour la formule du F de l'analyse de la variance. Au lieu de diviser par le carré moyen de l'erreur classique, ce carré moyen est ajusté aux variances de chaque groupe de l'ANOVA. L'interprétation de la p-valeur se fait de la même façon que pour le tableau d'analyse de la variance.

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.

Onglet **Général** :

Format des données :

Colonnes : activez cette option pour entrer les données sous forme de colonnes.

Tableau : activez cette option pour entrer les données sous forme de tableau. Un maximum de trois facteurs est alors autorisé.

Format des données = Colonnes :

- **Y / Variables dépendantes** :
    - **Quantitatives** : sélectionnez la ou les variables réponse que vous souhaitez modéliser. Si plusieurs variables sont sélectionnées, XLSTAT fera les calculs pour chacune des variables indépendamment. Si des en-têtes de colonnes ont été sélectionnés, veuillez vérifier que l'option « Libellés des variables » est activée.

- **X / Variables explicatives** :

- **Quantitatives** : activez cette option si vous voulez inclure une ou plusieurs variables explicatives quantitatives dans le modèle. Dans cas, vous ne ferez plus de l'ANOVA mais de l'ANCOVA. Sélectionnez alors la ou les variables correspondantes sur la feuille Excel. Les données sélectionnées doivent être de type numérique. Si le libellé des variables a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

- **Qualitatives** : sélectionnez la ou les variables explicatives qualitatives (les facteurs) sur la feuille Excel. Les données sélectionnées peuvent être de tout type, mais les données numériques sont automatiquement considérées comme nominales. Si le libellé des variables a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

Format des données = Tableau

- Tableau des données : Sélectionnez le tableau des données contenant la variable réponse et la ou les variables explicatives.
- Nombre de facteurs : Entrez le nombre de facteurs de votre analyse.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des variables : activez cette option si la première ligne des données sélectionnées (variables dépendantes et explicatives, libellés des observations, poids) contient un libellé.

Libellés des observations : activez cette option si vous voulez utiliser des libellés d'observations pour l'affichage des résultats. Si l'option « Libellés des variables » est activée, la première cellule de la sélection doit comprendre un en-tête. Si vous n'activez pas cette option, des libellés seront automatiquement créés (Obs1, Obs2, …).

Poids des observations : activez cette option si vous voulez pondérer les observations. Si vous n'activez pas cette option, les poids seront tous considérés comme valant 1. Les poids doivent être impérativement supérieurs ou égaux à 0. Un poids de 2 est équivalent à répéter deux fois la même observation. Si un en-tête de colonne a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

Poids dans la régression : activez cette option si vous voulez effectuer une régression par les moindres carrés pondérés. Si vous n'activez pas cette option, les poids seront tous considérés comme valant 1. Les poids doivent être impérativement supérieurs ou égaux à 0. Si un en-tête de colonne a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

Groupes : activez cette option puis sélectionnez ici les données d'appartenance à des groupes si vous souhaitez que les calculs soient effectués sur chaque groupe séparément.

Onglet **Options** :

Sous-onglet **Modèle** :

Constante fixée : activez cette option pour fixer la constante du modèle de régression à une valeur que vous devez ensuite saisir (0 par défaut).

Tolérance : activez cette option pour permettre à l'algorithme de calcul de la régression OLS ne pas prendre en compte les variables qui seraient soit constantes soit trop corrélées avec d'autres variables déjà utilisées dans le modèle (0.0001 par défaut).

Interactions / Niveau : activez cette option pour inclure des interactions dans le modèle puis entrez le niveau maximum d'interaction (valeur comprise entre 1 et 4).

Intervalle de confiance (%) : entrez l'étendue en pourcentage de l'intervalle de confiance à utiliser pour les différents tests, et pour le calcul des intervalles de confiance autour des paramètres et des prédictions. Valeur par défaut : 95.

Sélection du modèle : activez cette option si vous souhaitez utiliser l'une des quatre méthodes de sélection proposées :

- **Meilleur modèle** : cette méthode permet de choisir le meilleur modèle parmi tous les modèles comprenant un nombre de variables variant de « Min variables » à « Max variables ». Par ailleurs le « critère » pour déterminer le meilleur modèle peut être choisi par l'utilisateur.
- **Critère** : veuillez choisir le critère parmi la liste suivante : R^2 ajusté, Moyenne des Carrés des Erreurs (MCE), Cp de Mallows, AIC de Akaike, SBC de Schwarz, PC d'Amemiya.
- **Min variables** : entrez le nombre minimum de variables à prendre en compte dans le modèle.
- **Max variables** : entrez le nombre maximum de variables à prendre en compte dans le modèle.

Remarque : cette méthode peut entraîner des calculs longs car le nombre total de modèles explorés est la somme des Cn,k pour k variant entre « Min variables » et « Max variables », où Cn,k vaut n!/[(n-k)!k !]. Il est donc conseillé d'augmenter progressivement la valeur de « Max variables ».

- **Stepwise** : le processus de sélection commence par l'ajout de la variable ayant la plus forte contribution au modèle (le critère utilisé est la statistique t de Student). Si une seconde variable est telle que la probabilité associée à son t est inférieure à la « **Probabilité pour l'entrée** », elle est ajoutée au modèle. De même pour une troisième variable. A partir de l'ajout de la troisième variable, après chaque ajout, on évalue pour toutes les variables présentes dans le modèle quel serait l'impact de son retrait (toujours au travers de la statistique t). Si la probabilité est supérieure à la « **Probabilité pour le retrait** », la variable est retirée. La procédure se poursuit jusqu'à ce que plus aucune variable ne puisse être ajoutée/retirée.

- **Ascendante** : la procédure est identique à cette de la sélection progressive, hormis le fait que les variables sont uniquement ajoutées et jamais retirées.

- **Descendante** : la procédure commence par l'ajout simultané de toutes les variables. Les variables sont ensuite retirées du modèle suivant la procédure utilisée pour la sélection progressive.

Sous-onglet **ANOVA/ANCOVA** :

Contraintes : des détails sur les différentes options sont disponibles dans la section description.

- **a1 = 0** : choisissez cette option pour que le paramètre de la première modalité de chaque facteur soit fixé à 0.

- **an = 0** : choisissez cette option pour que le paramètre de la dernière modalité de chaque facteur soit fixé à 0.

- **Somme (ai) = 0** : pour chaque facteur la somme des paramètres associés aux différentes modalités vaut 0.

- **Somme (ni.ai) = 0** : pour chaque facteur la somme des paramètres associés aux différentes modalités pondérés par la fréquence des modalités respectives vaut 0.

Effets imbriqués : activez cette option pour inclure un effet imbriqué dans le modèle.

Effets aléatoires : activez cette option pour inclure des facteurs aléatoires dans le modèle. Leur impact ne se fera que sur les carrés moyens attendus de la table *Modèles mixtes - Analyse Type III Sum of Squares*.

Anova restreinte : activez cette option pour effectuer le calcul des effets mixtes (fixes/aléatoires) selon les contraintes imposées par l'anova restreinte. Leur impact ne se fera que sur les tests F de Fisher et les carrés moyens attendus de la table *Modèles mixtes - Analyse Type III Sum of Squares*.

Sous-onglet **Covariances** :

Dans cet onglet, vous pouvez choisir d'appliquer des corrections pour l'hétéroscedasticité et l'autocorrélation. Veuillez vous reporter à la section *Description* pour plus de détails.

Onglet **Validation** :

Validation : activez cette option si vous souhaitez utiliser une partie des données sélectionnées pour valider le modèle.

Jeu de validation : choisissez l'une des options pour définir le mode de sélection des observations utilisées pour la validation :

- **Aléatoire** : les observations sont sélectionnées de manière aléatoire. Le « Nombre d'observations » doit alors être saisi.
- **N dernières lignes** : les N dernières observations sont sélectionnées pour la validation. Le « Nombre d'observations » N doit alors être saisi.
- **N premières lignes** : les N premières observations sont sélectionnées pour la validation. Le « Nombre d'observations » N doit alors être saisi.
- **Variable de groupe** : si vous choisissez cette option, vous devez ensuite sélectionner une variable indicatrice composée de 0 pour les observations à utiliser pour le calcul du modèle, et de 1 pour les observations à utiliser pour la validation du modèle.

Onglet **Prédiction** :

Prédiction : activez cette option si vous souhaitez sélectionner des données à utiliser en mode prédiction. Si vous activez cette option, vous devez veiller à ce que les données de prédiction soient organisées comme les données d'estimation : mêmes variables, même ordre dans les sélections. En revanche vous ne devez pas sélectionner de libellés de variables : la première ligne des sélections décrites ci-dessous doit être une ligne de données.

Quantitatives : activez cette option pour sélectionner la ou les variables quantitatives explicatives. La première ligne ne doit pas comprendre d'en- tête.

Qualitatives : activez cette option pour sélectionner la ou les variables qualitatives explicatives. La première ligne ne doit pas comprendre d'en-tête.

Libellés des observations : activez cette option si vous voulez utiliser des libellés d'observations disponibles sur une feuille Excel pour l'affichage des résultats. La première ligne ne doit pas comprendre d'en-tête. Si vous n'activez pas cette option, des libellés seront automatiquement créés (PredObs1, PredObs2, …).

Onglet **Données manquantes** :

Supprimer les observations : activez cette option pour supprimer les observations comportant des données manquantes.

- **Vérifier pour chaque Y séparément** : choisissez cette option si vous voulez que lorsque, pour une observation donnée, il y a des données manquantes uniquement dans les Y, l'observation ne soit supprimée que si la donnée correspondant au Y en cours de modélisation est manquante.

- **Pour tous les Y** : choisissez cette option pour supprimer toutes les observations pour lesquelles des Y sont manquants.

- Remarque : les deux alternatives ci-dessus sont sans effet s'il n'y a qu'un seul Y.

Ignorer les données manquantes : si vous choisissez cette option, pour les données manquantes correspondant aux variables dépendantes XLSTAT essaiera de les estimer à partir du modèle obtenu. Pour celles correspondant aux variables explicatives, les observations correspondantes seront conservées dans la mesure du possible pour estimer la matrice de variance covariance (suppression par paire).

Estimer les données manquantes : activez cette option pour estimer les données manquantes avant le début des calculs.

- **Moyenne ou mode** : activez cette option pour estimer les données manquantes en utilisant la moyenne (variables quantitatives) ou le mode (variables qualitatives) pour les variables correspondantes.

- **Plus proche voisin** : activez cette option pour estimer les données manquantes d'une observation en recherchant le plus proche voisin de l'observation.

Onglet **Sorties** :

Sous-onglet **Général** :

Statistiques descriptives : activez cette option pour afficher les statistiques descriptives pour les variables sélectionnées.

Corrélations : activez cette option pour afficher la matrice de corrélation pour les variables quantitatives (dépendantes et explicatives).

Statistiques de multicolinéarité : activez cette option pour afficher les statistiques de multicolinéarité.

Mesures de la taille de l'effet : activez cette option pour afficher les mesures de la taille de l'effet. Dans le cadre de l'ANOVA, les mesures suivantes sont affichées : * éta carré : $\eta^2 = \frac{SC(Facteur)}{SC(Totale)}$ * éta carré partiel : $\eta_p^2 = \frac{SC(Facteur)}{SC(Facteur)+SC(Totale)}$ * omega carré : $\omega^2 = \frac{SC(Facteur)-ddl*MCE}{SC(Totale)+MCE}$ * F de Cohen : $f = \sqrt{\frac{\eta_p^2}{1-\eta_p^2}}$

Analyse de la variance : activez cette option pour afficher le tableau d'analyse de la variance.

Type I/II/III SS : activez cette option pour afficher les tableaux Type I SS, Type II SS et Type III SS permettant de mesurer la contribution des différentes variables explicatives au modèle (SS correspond à *Sum of Squares*).

Press : activez cette option pour calculer et afficher la statistique Press (predicted residual error sum of squares).

Interprétation : activez cette option pour que XLSTAT calcule une interprétation automatique des résultats.

Coefficients normalisés : activez cette option pour afficher les paramètres normalisés du modèle (coefficients bêta).

Prédictions et résidus : activez cette option pour afficher les prédictions et les résidus pour l'ensemble des observations.

- **Intervalles de confiance** : activez cette option pour calculer et afficher les intervalles de confiance sur les prédictions.
- **Prédictions ajustées** : activez cette option pour calculer et afficher les prédictions ajustées dans le tableau des prédictions et résidus.
- **Diagnostics d'influence** : activez cette option pour calculer et afficher le tableau des statistiques permettant d'identifier les observations ayant une influence sur les prédictions ou sur les coefficients associés à certaines variables explicatives (voir section résultats).

Tests de Welch et Brown-Forsythe : activez cette option pour afficher les tests de Welch et de Brown-Forsythe (voir la section description) dans le cas d'une ANOVA à un facteur.

Sous-onglet **Moyennes** :

Moyennes : activez cette option pour calculer et afficher les moyennes des modalités des variables qualitatives.

- **LS means** : activez cette option pour utiliser les LS means (Least square means) estimées à partir du modèle et non des observations.
- **Erreurs standard** : activez cette option pour calculer et afficher les erreurs standard associées aux moyennes.
- **Intervalles de confiance** : activez cette option pour calculer les intervalles de confiance autour des moyennes.
- **Trier** : activez cette option pour afficher les moyennes triées dans l'ordre croissant.
- **Comparaisons multiples** : activez cette option pour effectuer des tests de comparaisons multiples. Utilisez les LS means (Least square means) estimées à partir du modèle et non des observations. Des informations sur les tests de comparaisons multiples sont disponibles dans la section description.
- **Appliquer à tous les facteurs** : activez cette option pour calculer les tests sélectionnés pour tous les facteurs.

- **Trier en ordre croissant** : activez cette option pour trier les modalités comparées en ordre croissant, le critère de tri étant leur moyenne respective. Si cette option n'est pas activée, le tri est décroissant.
- **Comparaison par paires** : activez cette option puis choisissez les méthodes de comparaison.
- **Comparaison à un témoin** : activez cette option puis choisissez le type de test de Dunnett que vous voulez effectuer.
- **Choisir la MCE** : activez cette option pour choisir la variable dont l'erreur sera prise comme référence pour les comparaisons multiples. Dans le cadre de l'utilisation de modèles à facteurs aléatoires, l'utilisation de la moyenne des carrés des erreurs (MCE) associée au modèle complet (cas classique) n'est pas adaptée. On voudra choisir une moyenne des carrés des erreurs associée à un autre terme du modèle (en général un terme d'interaction). Si cette option est activée, une nouvelle boîte de dialogue vous permettant de sélectionner la variable à utiliser apparaît.
- **Protégé** : activez cette option pour empêcher que les tests de comparaisons multiples soient calculés si le facteur n'est pas significatif dans le modèle.
- **Boîte Top/Bottom** : activez cette option pour afficher les boîtes Top/Bottom (effectif des valeurs hautes et basses). Vous pouvez choisir entre Top/Bottom 2 ou 3. La boîte Top/Bottom 2 est l'effectif des deux valeurs les plus hautes/basses. La boîte Top/Bottom 3 est l'effectif des trois valeurs les plus hautes/basses.

Sous-onglet **Contrastes** :

Calculer contrastes : activer cette option pour calculer les contrastes, puis sélectionnez le tableau des contrastes, où il doit y avoir une colonne par contraste et une ligne pour chaque coefficient du modèle.

- **Correction de Bonferroni** : activez cette option appliquer la correction de Bonferroni pour prendre en compte le nombre de contrastes et ajuster le alpha.

Sous-onglet **Test des hypothèses** :

Ces options ne sont disponibles que si dans l'onglet Sorties/Général, l'option **prédictions et résidus** est activée.

Test de normalité : activez cette option pour qu'un test de Shapiro Wilk soit effectué sur les résidus.

Test de Levene : activez cette option pour qu'un test de Levene soit effectué afin de comparer les variances des différentes modalités pour chaque facteur.

Onglet **Graphiques** :

Options communes :

Graphiques de régression : activez cette option pour afficher les graphiques de régression :

- **Coefficients normalisés** : activez cette option pour afficher sur un graphique les paramètres normalisés du modèle avec leur intervalle de confiance.

- **Prédictions et résidus** : activez cette option pour afficher les graphiques suivants :

(1) Droite de régression : ce graphique n'est affiché que s'il n'y a qu'une seule variable explicative, et que cette variable est quantitative.

(2) Variable explicative versus résidus normalisés : ce graphique n'est affiché que s'il n'y a qu'une seule variable explicative, et que cette variable est quantitative.

(3) Variable dépendante versus résidus normalisés.

(4) Prédictions pour la variable dépendante versus variable dépendante.

(5) Graphique en bâtons des résidus normalisés.

- **Intervalles de confiance** : activez cette option pour afficher les intervalles de confiance sur les graphiques (1) et (4).

Graphiques des moyennes : activez cette option pour afficher les graphiques permettant de visualiser les moyennes pour les différentes modalités des différents facteurs.

Graphiques de synthèse : activez cette option pour afficher les graphiques permettant pour chaque facteur de comparer visuellement les moyennes pour les différentes modalités des différents facteurs et pour les différents variables dépendantes Y. Si l'option « filtrer les Y » est activée, ne sont présents sur chaque graphique que les résultats pour les Y pour lesquels le modèle est significatif.

Résultats

Statistiques descriptives : les tableaux de statistiques descriptives présentent pour toutes les variables sélectionnées des statistiques simples. Pour les variables dépendantes et les variables explicatives quantitatives, sont affichés le nombre d'observations, le nombre de données manquantes, le nombre de données non manquantes, la moyenne, et l'écart-type (non biaisé). Pour les variables explicatives qualitatives sont affichés le nom des différentes modalités ainsi que leur fréquence respective.

Matrice de corrélation : ce tableau est affiché afin de vous permettre d'avoir un aperçu des corrélations entre les différentes variables sélectionnées.

Synthèse de la sélection des variables : dans le cas où une méthode de sélection a été choisie, XLSTAT affiche la synthèse de la sélection. Dans le cas d'une sélection pas à pas, les statistiques correspondant aux différentes étapes sont affichées. Dans le cas d'une sélection du meilleur modèle pour un nombre de variables variant de p à q, le meilleur modèle pour chaque nombre de variable est affiché avec les statistiques correspondantes ; le meilleur modèle pour le critère choisi est alors affiché en gras.

Coefficients d'ajustement : dans ce tableau sont affichées les statistiques relatives à l'ajustement du modèle d'ANOVA :

- **Observations** : le nombre d'observations prises en compte dans les calculs. Dans les formules présentées ci-dessous n désigne le nombre d'observations.
- **Somme des poids** : la somme des poids des observations prises en compte dans les calculs. Dans les formules présentées ci-dessous W désigne la somme des poids.
- **DDL** : le nombre de degrés de liberté pour le modèle retenu (correspondant à la partie erreurs).
- **R²** : le coefficient de détermination du modèle. Sa valeur est comprise entre 0 et 1. Il est défini par :

$$R^2 = \frac{\sum_{i=1}^{n} w_i(\hat{y}_i - \bar{y}_i)^2}{\sum_{i=1}^{n} w_i(y_i - \bar{y}_i)^2} \text{ avec } \bar{y} = \frac{1}{W}\sum_{i=1}^{n} w_i y_i$$

Le R^2 s'interprète comme la proportion de la variabilité de la variable dépendante expliquée par le modèle. Plus le R^2 est proche de 1, meilleur est le modèle. L'inconvénient du R^2 est qu'il ne prend pas en compte le nombre de variables utilisées pour ajuster le modèle.

- **R² ajusté** : le coefficient de détermination ajusté du modèle. Le R^2 ajusté peut être négatif si le R^2 est voisin de zéro. Ce coefficient n'est affiché que si la constante du modèle n'est pas fixée par l'utilisateur. Il est défini par :

$$\hat{R}^2 = 1 - (1 - R^2)\frac{W - 1}{W - p - 1}$$

Le R^2 ajusté est une correction du R^2 qui permet de prendre en compte le nombre de variables utilisées dans le modèle.

- **MCE** : la moyenne des carrés des erreurs (MCE) est définie par :

$$MCE = \frac{1}{W - p^*}\sum_{i=1}^{n} w_i(y_i - \hat{y}_i)^2$$

- **RMCE** : la racine de la moyenne des carrés des erreurs (RMCE) est la racine carrée de la MCE.
- **MAPE** : la *Mean Absolute Percentage Error* est calculée comme suit :

$$MAPE = \frac{100}{W}\sum_{i=1}^{n} w_i \left|\frac{y_i - \hat{y}_i}{y_i}\right|$$

- **DW** : le coefficient de Durbin-Watson est défini par

$$DW = \frac{\sum_{i=2}^{n}\left[(y_i - \hat{y}_i) - (y_{i-1} - \hat{y}_{i-1})\right]^2}{\sum_{i=1}^{n} w_i(y_i - \hat{y}_i)^2}$$

Ce coefficient correspond au coefficient d'autocorrélation d'ordre 1 et permet de vérifier que les résidus du modèle ne sont pas autocorrélés, sachant que l'indépendance des résidus est l'une

des hypothèses de base de la régression linéaire. L'utilisateur pourra se référer à une table des coefficients de Durbin-Watson pour vérifier si l'hypothèse d'indépendance des résidus est acceptable.

- **Cp** : le coefficient Cp de Mallows est défini par

$$Cp = \frac{SCE}{\hat{\sigma}} + 2p^* - W$$

où SCE est la somme du carré des erreurs pour le modèle avec p variables explicatives, et où $\hat{\sigma}$ correspond à l'estimateur de la variance des résidus pour le modèle comprenant toutes les variables explicatives. Plus le coefficient Cp est proche de p* moins le modèle est biaisé.

- **AIC** : le critère d'information d'Akaike (Akaike's Information Criterion) est défini par

$$AIC = W \ln(\frac{SCE}{W}) + 2p^*$$

Ce critère proposé par Akaike (1973) dérive de la théorie de l'information, et s'appuie sur la mesure de Kullback et Leibler (1951). C'est un critère de sélection de modèles qui pénalise les modèles pour lesquels l'ajout de nouvelles variables explicatives n'apporte pas suffisamment d'information au modèle, l'information étant mesurée au travers de la SCE. On cherche à minimiser le critère AIC.

- **SBC** : le critère bayésien de Schwarz (Schwarz's Bayesian Criterion) est défini par

$$SBC = W \ln(\frac{SCE}{W}) + \ln(W)p^*$$

Ce critère proposé par Schwarz (1978) est proche du critère AIC, et comme ce dernier on cherche à le minimiser.

- **PC** : le critère de prédiction d'Amemiya (Amemiya's Prediction Criterion) est défini par

$$PC = \frac{(1-R^2)(W+p^*)}{W-p^*}$$

Ce critère proposé par Amemiya (1980) permet comme le R² ajusté de tenir compte de la parcimonie du modèle.

- **Press** : la statistique de Press n'est affichée que si l'option correspondante a été activée dans la boîte de dialogue. Elle est définie par

$$Press = \sum_{i=1}^{n} w_i \left(y_i - \hat{y}_{i(-i)}\right)^2$$

où $\hat{y}_{i(-i)}$ est la prédiction pour l'observation i lorsque cette dernière n'est pas utilisée pour l'estimation des paramètres. On obtient alors

$$\text{Press RMCE } = \sqrt{\frac{\text{Press}}{W-p^*}}$$

Le Press RMCE peut alors être comparé au RMCE. Une différence importante entre les deux indique que le modèle est sensible à la présence ou absence de certaines observations dans le modèle.

- **Q²** : cette statistique aussi connue comme le R² de validation croisée, n'est affichée que si l'option Press est activée. Elle est définie par :

$$Q^2 = 1 - \frac{\text{Press}}{\sum_{i=1}^{n}(y_i - \overline{y})^2}$$

Le Q² indique la proportion de la variance totale expliquée par les variables explicatives lorsque les prédictions pour chaque observation sont calculées lorsque l'observation en question n'est pas dans l'échantillon servant à l'estimation des paramètres. Une différence importante entre le Q² et le R² indique que le modèle est sensible à l'ajout ou au retrait de certaines observations dans l'échantillon d'estimation.

Le **tableau d'analyse de la variance** permet d'évaluer le pouvoir explicatif des variables explicatives. Dans le cas où la constante du modèle n'est pas fixée à une valeur donnée, le pouvoir explicatif est évalué en comparant l'ajustement (au sens des moindres carrés) du modèle final avec l'ajustement du modèle rudimentaire composé d'une constante égale à la moyenne de la variable dépendante. Dans le cas où la constante du modèle est fixée, la comparaison est faite par rapport au modèle pour lequel la variable dépendante serait égale à la constante fixée.

Si l'option Type I/III SS (SS : Sum of Squares) est activée, les tableaux suivants sont affichés.

Le tableau des **Type I SS** permet de visualiser l'influence de l'ajout progressif des variables explicatives sur l'ajustement du modèle, au sens de la somme des carrés des erreurs (SCE), de la moyenne des carrés des erreurs (MCE), du F de Fisher, ou de la probabilité associée au F de Fisher. Plus la probabilité est faible, plus la contribution de la variable au modèle est importante, toutes les autres variables étant déjà dans le modèle. Remarques : l'ordre de sélection des variables dans le modèle influe sur les valeurs obtenues ; la somme des sommes des carrés de ce tableau est égal à la somme des carrés du modèle.

Le tableau des **Type II SS** permet de visualiser l'influence du retrait d'une variable explicative sur l'ajustement du modèle, toutes les autres variables étant conservées, au sens de la somme des carrés des erreurs (SCE), de la moyenne des carrés des erreurs (MCE), du F de Fisher, ou de la probabilité associée au F de Fisher. Plus la probabilité est faible, plus la contribution de la variable au modèle est importante. Remarque : dans le cas des ANOVAs déséquilibrées, l'utilisation des Type III est recommandée mais XLSTAT affiche les Type II pour les utilisateurs avancés qui voudraient disposer des Type II.

Le tableau des **Type III SS** permet de visualiser l'influence du retrait d'une variable explicative sur l'ajustement du modèle, toutes les autres variables étant conservées, au sens de la somme des carrés des erreurs (SCE), de la moyenne des carrés des erreurs (MCE), du F de Fisher, ou de la probabilité associée au F de Fisher. Plus la probabilité est faible, plus la contribution de la variable au modèle est importante, toutes les autres variables étant déjà dans le modèle. Remarque : contrairement au cas des Type I SS, l'ordre de sélection des variables dans le modèle n'influe pas sur les valeurs obtenues, et contrairement aux Type II SS, les valeurs ne dépendent pas des effectifs des cellules (par cellule on entend une combinaison de modalités

des différents facteurs), ce qui fait des Type III le test recommandé pour évaluer la contribution d'une variable.

L'**équation du modèle** est ensuite affichée pour faciliter la lecture ou la réutilisation du modèle.

Le tableau des **coefficients normalisés** (aussi appelés coefficients bêta) permet de comparer le poids relatif des variables. Plus la valeur absolue d'un coefficient est élevée, plus le poids de la variable correspondante est important. Lorsque l'intervalle de confiance autour des coefficients normalisés comprend la valeur 0 (cela est facilement visible sur le graphique des coefficients normalisés), le poids d'une variable dans le modèle n'est pas significatif.

Dans le tableau des **prédictions et résidus** sont donnés pour chaque observation, son poids, la valeur de la variable explicative qualitative s'il n'y en a qu'une, la valeur observée de la variable dépendante, la prédiction du modèle, les résidus, les résidus studentisés, les intervalles de confiance, ainsi que la prédiction ajustée si l'option correspondante a été activée dans la boîte de dialogue. Deux types d'intervalles de confiance sont affichés : un intervalle de confiance autour de la moyenne (correspondant au cas où l'on ferait la prédiction pour un nombre infini d'observations avec un ensemble de valeurs données des variables explicatives) et un intervalle autour de la prédiction ponctuelle (correspondant au cas d'une prédiction isolée pour des valeurs données des variables explicatives). Le second intervalle est toujours plus grand que le premier, la variabilité étant plus importante. Si des données de validation ont été sélectionnées, elles sont affichées en fin de tableau.

Les **graphiques** qui suivent permettent de visualiser les résultats mentionnés ci-dessus. S'il n'y a qu'une seule variable explicative dans le modèle, le premier graphique affiché permet de visualiser les valeurs observées, la droite de régression et les deux types d'intervalles de confiance autour des prévisions. Le second graphique permet quant à lui de visualiser les résidus normalisés en fonction de la variable explicative. En principe, les résidus doivent être distribués de manière aléatoire autour de l'axe des abscisses. L'observation d'une tendance ou d'une forme révèlerait un problème au niveau du modèle.

Les **trois graphiques** affichés ensuite permettent de visualiser respectivement l'évolution des résidus normalisés en fonction de la variable dépendante, la distance entre les prédictions et les observations (pour un modèle idéal, les points seraient tous sur la bissectrice), et les résidus normalisés sur la forme d'un diagramme en bâtons. Ce dernier graphique permet de rapidement voir si un nombre anormal de données sort de l'intervalle]-2, 2[sachant que ce dernier, sous hypothèse de normalité, doit contenir environ 95% des données.

Dans le tableau des **diagnostics d'influence** sont affichés pour chaque observation, son poids, le résidu, le résidu normalisé (division par la RMCE), le résidu studentisé, le résidu supprimé (Deleted), le résidu supprimé studentisé, le leverage centré, la distance de Mahalanobis, le D de Cook, le CovRatio, le DFFits, le DFFits standardisé, les DFBeta (un par coefficient du modèle) et les DFBeta standardisés.

Quatre graphiques sont ensuite affichés pour mettre en évidence les observations dont l'influence nécessite une attention particulière.

Si vous avez sélectionné des données à utiliser pour calculer des **prédictions sur de nouvelles observations**, le tableau correspondant est ensuite affiché.

Si les tests pour la vérification des hypothèses de **normalité** et d'**homoscédasticité** ont été demandés, les résultats sont ensuite affichés.

Si des tests de **comparaisons multiples** ont été demandés, les résultats correspondants sont ensuite affichés.

Lorsqu'une ANOVA à un facteur a été appliquée et que l'option correspondante a été activée, les résultats des tests de Welch et de Brown-Forsythe sont affichés. On peut retrouver les statistiques associées, les degrés de libertés ainsi que les p-valeurs.

Si plusieurs variables dépendantes ont été sélectionnées et si l'option de comparaisons multiples a été activée, un tableau indiquant les moyennes de chaque modalité de chaque facteur et pour tous les Y est affiché. Les cellules du tableau sont colorées en utilisant une échelle spectrale allant du bleu au rouge. S'il y a plus de 10 catégories, seules les 5 moyennes les plus faibles et 5 plus élevées sont colorées. Un graphique permet de visualiser les mêmes résultats.

Un deuxième graphique permet de visualiser les moyennes estimées accompagnées des lettres de regroupement issues des tableaux de comparaisons multiples. Ce deuxième graphique n'affiche les résultats que pour les variables dépendantes significatives pour lesquelles le test F est significatif.

Exemple

Un exemple d'ANOVA à un facteur est disponible sur le Centre d'aide XLSTAT à l'adresse suivante :

http://www.xlstat.com/demo-anof.htm

Un exemple d'ANOVA à deux facteurs avec interaction est disponible sur le Centre d'aide XLSTAT à l'adresse suivante :

http://www.xlstat.com/demo-ano2f.htm

Bibliographie

Akaike H. (1973). Information theory and the extension of the maximum likelihood principle. In: Second International Symposium on Information Theory. (Eds: V.N. Petrov and F. Csaki). Academiai Kiadó, Budapest. 267-281.

Amemiya T. (1980). Selection of regressors. *International Economic Review*, **21**, 331-354.

Brown M. B. and Forsythe A. B. (1974). The ANOVA and multiple comparisons for data with heterogeneous variances. *Biometrics*, **30**, 719-724.

Dempster A.P. (1969). Elements of Continuous Multivariate Analysis. Addison-Wesley, Reading.

Hsu J.C. (1996). Multiple Comparisons: Theory and Methods. CRC Press, Boca Raton.

Jobson J. D. (1999). Applied Multivariate Data Analysis: Volume 1: Regression and Experimental Design. Springer Verlag, New York.

Lea P., Naes T. and Robotten M. (1997). Analysis of Variance for Sensory Data. John Wiley and Sons, London.

Mallows C.L. (1973). Some comments on Cp. *Technometrics*, **15**, 661-675.

Rogers W. H. (1993). Regression standard errors in clustered samples. _Stata Technical _Bulletin, **13**, 19–23.

Sahai H. and Ageel M.I. (2000). The Analysis of Variance. Birkhaüser, Boston.

Tomassone R., Audrain S., Lesquoy de Turckheim E. and Miller C. (1992). La Régression, Nouveaux Regards sur une Ancienne Méthode Statistique. INRA et MASSON, Paris.

Welch, B. L. (1951). On the comparison of several mean values: An alternative approach. *Biometrika*, **38**, 330-336.

Velleman P.F. and R.E. Welsch (1981). Efficient computing of regression diagnostics. *The American Statistician*, **35**, 234-242.

Welch B. L. (1951). On the comparison of several mean values: An alternative approach. *Biometrika*, **38**, 330-336.

Welsch R.E. and Kuh E. (1977). Linear Regression Diagnostics. *Sloan School of Management Working Paper*,923-977, M.I.T., Cambridge, Mass.

White H. (1980). A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity. *Econometrica*, **48(4)**, 817-838.

Zeileis A. (2006). Object-Oriented Computation of Sandwich Estimators. *Journal of Statistical Software*,**16(9)**, 1-16.

ANCOVA

Utilisez ce module pour modéliser une variable dépendante quantitative en utilisant des variables explicatives quantitatives et qualitatives dans le cadre du modèle linéaire.

Dans cette section :

Description

L'ANCOVA (Analyse de covariance) peut être vue comme un mélange d'ANOVA et de régression linéaire puisque la variable dépendante est de même nature, le modèle est aussi un modèle linéaire, et les hypothèses sont identiques. Il en réalité est plus juste de considérer l'ANOVA et la régression linéaire comme des cas particuliers de l'ANCOVA.

Si p est le nombre de variables quantitatives et q est le nombre de facteurs (les variables qualitatives, y compris les interactions entre variables qualitatives), le modèle de l'ANCOVA s'écrit de la manière suivante :

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \sum_{j=1}^{q} \beta_{k(i,j),j} + \epsilon_i \tag{1}$$

où y_i est la valeur observée pour la variable dépendante pour l'observation i, x_{ij} est la valeur prise par la variable quantitative j pour l'observation i, $k(i,j)$ est l'indice correspondant à la modalité du facteur j pour l'observation i, et ϵ_i est l'erreur du modèle.

Les hypothèses utilisées en ANCOVA sont identiques à celles de la régression linéaire et de l'ANOVA : les erreurs ϵ_i suivent une même loi normale $N(0, s)$ et sont indépendantes.

Les estimateurs des coefficients β et de leur variance sont donnés par :

$$\hat{\beta} = (X^t X)^{-1} X^t Y \tag{2}$$

et

$$Var(\hat{\beta}) = \hat{\sigma}^2 (X^t X)^{-1} \tag{3}$$

Si l'on souhaite utiliser les différents tests proposés dans les résultats de la régression linéaire il est recommandé de vérifier a posteriori que les hypothèses sous-jacentes sont bien vérifiées.

La normalité des résidus peut être vérifiée en analysant certains graphiques ou en utilisant un test de normalité. L'indépendance des résidus peut être vérifiée en analysant certains graphiques ou en utilisant le test de Durbin Watson disponible dans les outils d'analyse de séries chronologiques de XLSTAT. L'homoscédasticité peut être vérifiée au travers d'un test de Levene.

Interactions entre variables quantitatives et facteurs

L'une des spécificités de l'ANCOVA est de permettre la prise en compte d'interactions entre les variables quantitatives et les facteurs. La principale application est de permettre de tester si le niveau d'un facteur (une variable qualitative) a une influence sur le coefficient (souvent appelé pente dans ce contexte) d'une variable quantitative. Des tests de comparaison permettent alors tester si les pentes correspondant aux différents niveaux d'un facteur sont significativement différentes ou non. Un modèle à une variable quantitative et un facteur avec interaction s'écrit

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_{k(i,1),1} + \beta_{k(i,1),2} x_{i1} + \epsilon_i . \tag{4}$$

On peut simplifier cette écriture en posant

$$\gamma_{k(i,1),1} = \beta_1 + \beta_{k(i,1),2} \tag{5}$$

d'où l'on tire

$$y_i = \beta_0 + \beta_{k(i,1),1} + \gamma_{k(i,1),1} x_{i1} + \epsilon_i . \tag{6}$$

La comparaison des paramètres γ permet de tester si le facteur a un effet sur la pente.

Correction de l'hétéroscédasticité et de l'autocorrélation

L'homoscedasticité et l'indépendance des résidus (termes d'erreur) sont des hypothèses clefs de la régression, où, pour rappel, on suppose qu'ils sont indépendants, identiquement distribués suivant une loi normale de moyenne nulle. Lorsque ces hypothèses ne peuvent être validées (un test de Durbin Watson ou de White disponibles dans les outils d'analyse des séries chronologiques permettent de les vérifier), une conséquence est que la matrice de covariance ne peut-être calculée suivant la formule (2). Les variances des coefficients *b* peuvent alors être fausses et les conclusions quant à la significativité de la contribution ou non au modèle des variables correspondantes peuvent alors être faussées, de même que les intervalles de confiance. Une variable explicative pourrait être déclarée comme inutile alors que sa contribution est significative. XLSTAT permet de corriger les matrices de covariance pour les effets d'hétéroscédasticité et d'autocorrélation qui peuvent survenir notamment dans des cas où le temps intervient (séries chronologiques, données longitudinales).

Pour ce qui concerne l'hétéroscédasticité, White (1980) suivi par plusieurs auteurs, a exploré plusieurs façons d'améliorer le calcul de la matrice de variance-covariance des paramètres β , en prenant en compte les résidus et les *leverages* centrés obtenus à partir des calculs standards de la régression linéaire (voir MacKinnon (1985) et Zeileis (2006) pour une revue exhaustive). Lorsque les hypothèses de la régression linéaire ne peuvent être conservées, si les estimateurs des coefficients ne sont pas modifiés, la simplification permettant d'aboutir à l'équation (3) n'est plus possible, et l'on doit revenir à l'expression générale suivante :

$$Var(\beta) = (X^t X)^{-1} (X^t \Omega X) (X^t X)^{-1} . \tag{7}$$

L'équation (7) est équivalente à l'équation (3) lorsque

$$\Omega = \hat{\sigma}^2 I. \tag{8}$$

Soient ω_i les éléments de la diagonale de W. Les différents estimateurs consistents pour l'hétéroscédasticité proposés (HC, heteroscedasticity consistent) pour les ω_i sont donnés par :

$$\begin{aligned}
HC0: &\quad \omega_i = {\hat{e_i}}^2 \\
HC1: &\quad \omega_i = {\hat{e_i}}^2 \frac{n}{(n-p-1)} \\
HC2: &\quad \omega_i = \frac{{\hat{e_i}}^2}{(1-h_i)} \\
HC3: &\quad \omega_i = \frac{{\hat{e_i}}^2}{(1-h_i)^2} \\
HC4: &\quad \omega_i = \frac{{\hat{e_i}}^2}{(1-h_i)^{\delta_i}} \text{ avec } \delta_i = min(4, h_i/\bar{h})
\end{aligned}$$

Où les $\hat{e_i}$ sont les résidus, et les h_i sont les *leverages* centrés, et p est le nombre de variables explicatives.

Newey et West (1987) ont suggéré un estimateur qui permet d'appliquer une correction à la fois pour l'autocorrélation et pour l'hétéroscédasticité, mais le décalage (lag) doit être connu de l'utilisateur (les outils d'analyse descriptive des séries chronologiques ou ARIMA de XLSTAT peuvent être utilisés pour cela). Pour un décalage de 0 (pas d'autocorrélation) nous avons :

$$X^t \Omega X = X^t \Omega_0 X = \frac{n}{n-p-1} \sum_{i=1}^{n} \hat{e}_i^2 x_i^t x_i$$

où x_i est le vecteur des variables explicatives (incluant un 1 pour l'intercept du modèle) pour la ième observation. Pour un décalage de m pas (m>0), nous avons :

$$X^t \Omega X = X^t \Omega_0 X + \frac{n}{n-p-1} \sum_{l=1}^{m} \sum_{t=l+1}^{n} \hat{e}_t^2 \hat{e}_{t-l}^2 (x_t^t x_{t-l} - x_{t-l}^t x_t)\cdot$$

La version inajustée de l'estimateur de Newey et West correspond à la même approche sans le facteur de correction $n/(n-p-1)$.

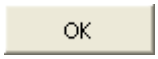
L'option **Classes** permet de corriger le problème d'hétéroscedasticité dans le cas où l'on considère que les variances sont égales uniquement à l'intérieur de groupes donnés. Lorsque cette option est sélectionnée, vous devez ensuite sélectionner les données indiquant à quel classe appartient chaque observation.

$$X^t \Omega X = \frac{n-1}{n-p-1} \frac{K}{K-1} \sum_{g=1}^{K} X_g^t \hat{\mathbf{e}} \hat{\mathbf{e}}^{\mathbf{t}} X_g$$

où K est le nombre de classes et X_g est le sous-ensemble d'observations correspondant à la classe g.

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.

Onglet **Général**:

Y / Variables dépendantes :

Quantitatives : sélectionnez la ou les variables réponse que vous souhaitez modéliser. Si plusieurs variables sont sélectionnées, XLSTAT fera les calculs pour chacune des variables indépendamment. Si des en-têtes de colonnes ont été sélectionnés, veuillez vérifier que l'option « Libellés des variables » est activée.

X / Variables explicatives :

Quantitatives : sélectionnez la ou les variables explicatives quantitatives sur la feuille Excel. Les données sélectionnées doivent être de type numérique. Si le libellé des variables a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

Qualitatives : sélectionnez la ou les variables explicatives qualitatives (les facteurs) sur la feuille Excel. Les données sélectionnées peuvent être de tout type, mais les données numériques sont automatiquement considérées comme nominales. Si le libellé des variables a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des variables : activez cette option si la première ligne des données sélectionnées (variables dépendantes et explicatives, libellés des observations, poids) contient un libellé.

Libellés des observations : activez cette option si vous voulez utiliser des libellés d'observations pour l'affichage des résultats. Si l'option « Libellés des variables » est activée, la première cellule de la sélection doit comprendre un en-tête. Si vous n'activez pas cette option, des libellés seront automatiquement créés (Obs1, Obs2, …).

Poids des observations : activez cette option si vous voulez pondérer les observations. Si vous n'activez pas cette option, les poids seront tous considérés comme valant 1. Les poids doivent être impérativement supérieurs ou égaux à 0. Un poids de 2 est équivalent à répéter deux fois la même observation. Si un en-tête de colonne a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

Poids dans la régression : activez cette option si vous voulez effectuer une régression par les moindres carrés pondérés. Si vous n'activez pas cette option, les poids seront tous considérés comme valant 1. Les poids doivent être impérativement supérieurs ou égaux à 0. Si un en-tête de colonne a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

Groupes : activez cette option puis sélectionnez ici les données d'appartenance à des groupes si vous souhaitez que les calculs soient effectués sur chaque groupe séparément.

Onglet **Options**:

Sous-onglet **Modèle**:

Constante fixée : activez cette option pour fixer la constante du modèle de régression à une valeur que vous devez ensuite saisir (0 par défaut).

Tolérance : activez cette option pour permettre à l'algorithme de calcul de la régression OLS ne pas prendre en compte les variables qui seraient soit constantes soit trop corrélées avec d'autres variables déjà utilisées dans le modèle (0.0001 par défaut).

Interactions / Niveau : activez cette option pour inclure des interactions dans le modèle puis entrez le niveau maximum d'interaction (valeur comprise entre 1 et 4).

Intervalle de confiance (%) : entrez l'étendue en pourcentage de l'intervalle de confiance à utiliser pour les différents tests, et pour le calcul des intervalles de confiance autour des paramètres et des prédictions. Valeur par défaut : 95.

Sélection du modèle : activez cette option si vous souhaitez utiliser l'une des quatre méthodes de sélection proposées :

- **Meilleur modèle** : cette méthode permet de choisir le meilleur modèle parmi tous les modèles comprenant un nombre de variables variant de « Min variables » à « Max variables ». Par ailleurs le « critère » pour déterminer le meilleur modèle peut être choisi par l'utilisateur.

- **Critère** : veuillez choisir le critère parmi la liste suivante : R² ajusté, Moyenne des Carrés des Erreurs (MCE), Cp de Mallows, AIC de Akaike, SBC de Schwarz, PC d'Amemiya.

- **Min variables** : entrez le nombre minimum de variables à prendre en compte dans le modèle.

- **Max variables** : entrez le nombre maximum de variables à prendre en compte dans le modèle.

Remarque : cette méthode peut entraîner des calculs longs car le nombre total de modèles explorés est la somme des $C_{n,k}$ pour k variant entre « Min variables » et « Max variables », où $C_{n,k}$ vaut $n!/[(n-k)!k!]$. Il est donc conseillé d'augmenter progressivement la valeur de « Max variables ».

- **Stepwise** : le processus de sélection commence par l'ajout de la variable ayant la plus forte contribution au modèle (le critère utilisé est la statistique t de Student). Si une seconde variable est telle que la probabilité associée à son t est inférieure à la « **Probabilité pour l'entrée** », elle est ajoutée au modèle. De même pour une troisième variable. A partir de l'ajout de la troisième variable, après chaque ajout, on évalue pour toutes les variables présentes dans le modèle quel serait l'impact de son retrait (toujours au travers de la statistique t). Si la probabilité est supérieure à la « **Probabilité pour le retrait** », la variable est retirée. La procédure se poursuit jusqu'à ce que plus aucune variable ne puisse être ajoutée/retirée.

- **Ascendante** : la procédure est identique à cette de la sélection progressive, hormis le fait que les variables sont uniquement ajoutées et jamais retirées.

- **Descendante** : la procédure commence par l'ajout simultané de toutes les variables. Les variables sont ensuite retirées du modèle suivant la procédure utilisée pour la sélection progressive.

Sous-onglet **ANOVA/ANCOVA**:

Contraintes : des détails sur les différentes options sont disponibles dans la section description.

- **a1 = 0** : choisissez cette option pour que le paramètre de la première modalité de chaque facteur soit fixé à 0.

- **an = 0** : choisissez cette option pour que le paramètre de la dernière modalité de chaque facteur soit fixé à 0.

- **Somme (ai) = 0** : pour chaque facteur la somme des paramètres associés aux différentes modalités vaut 0.
- **Somme (ni.ai) = 0** : pour chaque facteur la somme des paramètres associés aux différentes modalités pondérés par la fréquence des modalités respectives vaut 0.

Effets imbriqués : activez cette option pour inclure un effet imbriqué dans le modèle.

Effets aléatoires : activez cette option pour inclure des facteurs aléatoires dans le modèle. Leur impact ne se fera que sur les carrés moyens attendus.

Onglet **Validation** :

Validation : activez cette option si vous souhaitez utiliser une partie des données sélectionnées pour valider le modèle.

Jeu de validation : choisissez l'une des options pour définir le mode de sélection des observations utilisées pour la validation :

- **Aléatoire** : les observations sont sélectionnées de manière aléatoire. Le « Nombre d'observations » doit alors être saisi.
- **N dernières lignes** : les N dernières observations sont sélectionnées pour la validation. Le « Nombre d'observations » N doit alors être saisi.
- **N premières lignes** : les N premières observations sont sélectionnées pour la validation. Le « Nombre d'observations » N doit alors être saisi.
- **Variable de groupe** : si vous choisissez cette option, vous devez ensuite sélectionner une variable indicatrice composée de 0 pour les observations à utiliser pour le calcul du modèle, et de 1 pour les observations à utiliser pour la validation du modèle.

Onglet **Prédiction** :

Prédiction : activez cette option si vous souhaitez sélectionner des données à utiliser en mode prédiction. Si vous activez cette option, vous devez veiller à ce que les données de prédiction soient organisées comme les données d'estimation : mêmes variables, même ordre dans les sélections. En revanche vous ne devez pas sélectionner de libellés de variables : la première ligne des sélections décrites ci-dessous doit être une ligne de données.

Quantitatives : activez cette option pour sélectionner la ou les variables quantitatives explicatives. La première ligne ne doit pas comprendre d'en- tête.

Qualitatives : activez cette option pour sélectionner la ou les variables qualitatives explicatives. La première ligne ne doit pas comprendre d'en-tête.

Libellés des observations : activez cette option si vous voulez utiliser des libellés d'observations disponibles sur une feuille Excel pour l'affichage des résultats. La première ligne ne doit pas comprendre d'en-tête. Si vous n'activez pas cette option, des libellés seront automatiquement créés (PredObs1, PredObs2, …).

Onglet **Données manquantes** :

Supprimer les observations : activez cette option pour supprimer les observations comportant des données manquantes.

- **Vérifier pour chaque Y séparément** : choisissez cette option si vous voulez que lorsque, pour une observation donnée, il y a des données manquantes uniquement dans les Y, l'observation ne soit supprimée que si la donnée correspondant au Y en cours de modélisation est manquante.
- **Pour tous les Y** : choisissez cette option pour supprimer toutes les observations pour lesquelles des Y sont manquants.
- Remarque : les deux alternatives ci-dessus sont sans effet si il n'y a qu'un seul Y.

Ignorer les données manquantes : si vous choisissez cette option, pour les données manquantes correspondant aux variables dépendantes XLSTAT essaiera de les estimer à partir du modèle obtenu. Pour celles correspondant aux variables explicatives, les observations correspondantes seront conservées dans la mesure du possible pour estimer la matrice de variance covariance (suppression par paire).

Estimer les données manquantes : activez cette option pour estimer les données manquantes avant le début des calculs.

- **Moyenne ou mode** : activez cette option pour estimer les données manquantes en utilisant la moyenne (variables quantitatives) ou le mode (variables qualitatives) pour les variables correspondantes.
- **Plus proche voisin** : activez cette option pour estimer les données manquantes d'une observation en recherchant le plus proche voisin de l'observation.

Onglet **Sorties**:

Sous-onglet **Général**:

Statistiques descriptives : activez cette option pour afficher les statistiques descriptives pour les variables sélectionnées.

Corrélations : activez cette option pour afficher la matrice de corrélation pour les variables quantitatives (dépendantes et explicatives).

Statistiques de multicolinéarité : activez cette option pour afficher les statistiques de multicolinéarité.

Analyse de la variance : activez cette option pour afficher le tableau d'analyse de la variance.

Type I/II/III SS : activez cette option pour afficher les tableaux Type I SS, Type II SS et Type III SS permettant de mesurer la contribution des différentes variables explicatives au modèle (SS correspond à *Sum of Squares*).

Press : activez cette option pour calculer et afficher la statistique Press (predicted residual error sum of squares).

Interprétation : activez cette option pour que XLSTAT calcule une interprétation automatique des résultats.

Coefficients normalisés : activez cette option pour afficher les paramètres normalisés du modèle (coefficients bêta).

Prédictions et résidus : activez cette option pour afficher les prédictions et les résidus pour l'ensemble des observations.

- **X** : activez cette option pour afficher dans le tableau des prédictions, pour chaque observation, les données correspondant aux différentes variables explicatives quantitatives.
- **Intervalles de confiance** : activez cette option pour calculer et afficher les intervalles de confiance sur les prédictions.
- **Prédictions ajustées** : activez cette option pour calculer et afficher les prédictions ajustées dans le tableau des prédictions et résidus.
- **Diagnostics d'influence** : activez cette option pour calculer et afficher le tableau des statistiques permettant d'identifier les observations ayant une influence sur les prédictions ou sur les coefficients associés à certaines variables explicatives (voir section résultats).

Tests de Welch et Brown-Forsythe : activez cette option pour afficher les tests de Welch et de Brown-Forsythe (voir la section description) dans le cas d'une ANOVA à un facteur.

Sous-onglet **Moyennes** :

Moyennes : activez cette option pour calculer et afficher les moyennes des modalités des variables qualitatives.

- **LS means** : activez cette option pour utiliser les LS means (Least square means) estimées à partir du modèle et non des observations.
- **Erreurs standard** : activez cette option pour calculer et afficher les erreurs standard associées aux moyennes.
- **Intervalles de confiance** : activez cette option pour calculer les intervalles de confiance autour des moyennes.
- **Trier** : activez cette option pour afficher les moyennes triées dans l'ordre croissant.
- **Comparaisons multiples** : activez cette option pour effectuer des tests de comparaisons multiples.utiliser les LS means (Least square means) estimées à partir du modèle et non des observations. Des informations sur les tests de comparaisons multiples sont disponibles dans la section description de l'ANOVA.

- **Appliquer à tous les facteurs** : activez cette option pour calculer les tests sélectionnés pour tous les facteurs.
- **Trier en ordre croissant** : activez cette option pour trier les modalités comparées en ordre croissant, le critère de tri étant leur moyenne respective. Si cette option n'est pas activée, le tri est décroissant.
- **Comparaison par paires** : activez cette option puis choisissez les méthodes de comparaison.
- **Comparaison à un témoin** : activez cette option puis choisissez le type de test de Dunnett que vous voulez effectuer.
- **Choisir la MCE** : activez cette option pour choisir la variable dont l'erreur sera prise comme référence pour les comparaisons multiples. Dans le cadre de l'utilisation de modèles à facteurs aléatoires, l'utilisation de la moyenne des carrés des erreurs (MCE) associée au modèle complet (cas classique) n'est pas adaptée. On voudra choisir une moyenne des carrés des erreurs associée à un autre terme du modèle (en général un terme d'interaction). Si cette option est activée, une nouvelle boîte de dialogue vous permettant de sélectionner la variable à utiliser apparaît.
- **Protégé** : activez cette option pour empêcher que les tests de comparaisons multiples soient calculés si le facteur n'est pas significatif dans le modèle.
- **Boîte Top/Bottom** : activez cette option pour afficher les boîtes Top/Bottom (effectif des valeurs hautes et basses). Vous pouvez choisir entre Top/Bottom 2 ou 3. La boîte Top/Bottom 2 est l'effectif des deux valeurs les plus hautes/basses. La boîte Top/Bottom 3 est l'effectif des deux valeurs les plus hautes/basses.
- **Comparaison des pentes** : activez cette option pour comparer les pentes des interactions entre les variables quantitatives et qualitatives (voir la section description sur ce sujet).

Sous-onglet **Contrastes** :

Calculer contrastes : activer cette option pour calculer les contrastes, puis sélectionnez le tableau des contrastes, où il doit y avoir une colonne par contraste et une ligne pour chaque coefficient du modèle.

Sous-onglet **Test des hypothèses** :

Ces options ne sont disponibles que si dans l'onglet Sorties/Général, l'option **prédictions et résidus** est activée.

Test de normalité : activez cette option pour qu'un test de Shapiro Wilk soit effectué sur les résidus.

Test de Levene : activez cette option pour qu'un test de Levene soit effectué afin de comparer les variances des différentes modalités pour cha Shapiro Wilk soit effectué sur les résidus.

Onglet **Graphiques** :

Options communes :

Graphiques de régression : activez cette option pour afficher les graphiques de régression :

- **Coefficients normalisés** : activez cette option pour afficher sur un graphique les paramètres normalisés du modèle avec leur intervalle de confiance.
- **Prédictions et résidus** : activez cette option pour afficher les graphiques suivants :

(1) Droite de régression : ce graphique n'est affiché que s'il n'y a qu'une seule variable explicative, et que cette variable est quantitative.

(2) Variable explicative versus résidus normalisés : ce graphique n'est affiché que s'il n'y a qu'une seule variable explicative, et que cette variable est quantitative.

(3) Variable dépendante versus résidus normalisés.

(4) Prédictions pour la variable dépendante versus variable dépendante.

(5) Graphique en bâtons des résidus normalisés.

- **Intervalles de confiance** : activez cette option pour afficher les intervalles de confiance sur les graphiques (1) et (4).

Graphiques des moyennes : activez cette option pour afficher les graphiques permettant de visualiser les moyennes pour les différentes modalités des différents facteurs.

Graphiques de sytnhèse : activez cette option pour afficher les graphiques permettant pour chaque facteur de comparer visuellement les moyennes pour les différentes modalités des différents facteurs et pour les différents variables dépendantes Y. Si l'option « filtrer les Y est activée », ne sont présents sur chaque graphique que les résultats pour les Y pour lesquels le modèle est significatif.

Résultats

Statistiques descriptives : les tableaux de statistiques descriptives présentent pour toutes les variables sélectionnées des statistiques simples. Pour les variables dépendantes et les variables explicatives quantitatives, sont affichés le nombre d'observations, le nombre de données manquantes, le nombre de données non manquantes, la moyenne, et l'écart-type (non biaisé). Pour les variables explicatives qualitatives sont affichés le nom des différentes modalités ainsi que leur fréquence respective.

Matrice de corrélation : ce tableau est affiché afin de vous permettre d'avoir un aperçu des corrélations entre les différentes variables sélectionnées.

Synthèse de la sélection des variables : dans le cas où une méthode de sélection a été choisie, XLSTAT affiche la synthèse de la sélection. Dans le cas d'une sélection pas à pas, les statistiques correspondant aux différentes étapes sont affichées. Dans le cas d'une sélection du meilleur modèle pour un nombre de variables variant de p à q, le meilleur modèle pour chaque nombre de variable est affiché avec les statistiques correspondantes ; le meilleur modèle pour le critère choisi est alors affiché en gras.

Coefficients d'ajustement : dans ce tableau sont affichées les statistiques relatives à l'ajustement du modèle d'ANCOVA :

- **Observations** : le nombre d'observations prises en compte dans les calculs. Dans les formules présentées ci-dessous n désigne le nombre d'observations.
- **Somme des poids** : la somme des poids des observations prises en compte dans les calculs. Dans les formules présentées ci-dessous W désigne la somme des poids.
- **DDL** : le nombre de degrés de liberté pour le modèle retenu (correspondant à la partie erreurs).
- **R²** : le coefficient de détermination du modèle. Sa valeur est comprise entre 0 et 1. Il est défini par :

$$R^2 = \frac{\sum_{i=1}^{n} w_i(\hat{y}_i - \bar{y}_i)^2}{\sum_{i=1}^{n} w_i(y_i - \bar{y}_i)^2} \text{ avec } \bar{y} = \frac{1}{W}\sum_{i=1}^{n} w_i y_i$$

Le R^2 s'interprète comme la proportion de la variabilité de la variable dépendante expliquée par le modèle. Plus le R^2 est proche de 1, meilleur est le modèle. L'inconvénient du R^2 est qu'il ne prend pas en compte le nombre de variables utilisées pour ajuster le modèle.

- **R² ajusté** : le coefficient de détermination ajusté du modèle. Le R^2 ajusté peut être négatif si le R^2 est voisin de zéro. Ce coefficient n'est affiché que si la constante du modèle n'est pas fixée par l'utilisateur. Il est défini par :

$$\hat{R}^2 = 1 - (1 - R^2)\frac{W - 1}{W - p - 1}.$$

Le R^2 ajusté est une correction du R^2 qui permet de prendre en compte le nombre de variables utilisées dans le modèle.

- **MCE** : la moyenne des carrés des erreurs (MCE) est définie par :

$$MCE = \frac{1}{W - p^*}\sum_{i=1}^{n} w_i(y_i - \hat{y}_i)^2.$$

- **RMCE** : la racine de la moyenne des carrés des erreurs (RMCE) est la racine carrée de la MCE.
- **MAPE** : la *Mean Absolute Percentage Error* est calculée comme suit :

$$MAPE = \frac{100}{W}\sum_{i=1}^{n} w_i \left|\frac{y_i - \hat{y}_i}{y_i}\right|.$$

- **DW** : le coefficient de Durbin-Watson est défini par

$$DW = \frac{\sum_{i=2}^{n}\left[(y_i - \hat{y}_i) - (y_{i-1} - \hat{y}_{i-1})\right]^2}{\sum_{i=1}^{n} w_i(y_i - \hat{y}_i)^2}.$$

Ce coefficient correspond au coefficient d'autocorrélation d'ordre 1 et permet de vérifier que les résidus du modèle ne sont pas autocorrélés, sachant que l'indépendance des résidus est l'une des hypothèses de base de la régression linéaire. L'utilisateur pourra se référer à une table des coefficients de Durbin-Watson pour vérifier si l'hypothèse d'indépendance des résidus est acceptable.

- **Cp** : le coefficient Cp de Mallows est défini par

$$Cp = \frac{SCE}{\hat{\sigma}} + 2p^* - W\cdot$$

où SCE est la somme du carré des erreurs pour le modèle avec p variables explicatives, et où $\hat{\sigma}$ correspond à l'estimateur de la variance des résidus pour le modèle comprenant toutes les variables explicatives. Plus le coefficient Cp est proche de p* moins le modèle est biaisé.

- **AIC** : le critère d'information d'Akaike (Akaike's Information Criterion) est défini par

$$AIC = W\ln(\frac{SCE}{W}) + 2p^*\cdot$$

Ce critère proposé par Akaike (1973) dérive de la théorie de l'information, et s'appuie sur la mesure de Kullback et Leibler (1951). C'est un critère de sélection de modèles qui pénalise les modèles pour lesquels l'ajout de nouvelles variables explicatives n'apporte pas suffisamment d'information au modèle, l'information étant mesurée au travers de la SCE. On cherche à minimiser le critère AIC.

- **SBC** : le critère bayésien de Schwarz (Schwarz's Bayesian Criterion) est défini par

$$SBC = W\ln(\frac{SCE}{W}) + \ln(W)p^*\cdot$$

Ce critère proposé par Schwarz (1978) est proche du critère AIC, et comme ce dernier on cherche à le minimiser.

- **PC** : le critère de prédiction d'Amemiya (Amemiya's Prediction Criterion) est défini par

$$PC = \frac{(1-R^2)(W+p^*)}{W-p^*}.$$

Ce critère proposé par Amemiya (1980) permet comme le R² ajusté de tenir compte de la parcimonie du modèle.

- **Press** : la statistique de Press n'est affichée que si l'option correspondante a été activée dans la boîte de dialogue. Elle est définie par

$$Press = \sum_{i=1}^{n} w_i(y_i - \hat{y}_{i(-i)})^2,$$

où $\hat{y}_{i(-i)}$ est la prédiction pour l'observation i lorsque cette dernière n'est pas utilisée pour l'estimation des paramètres. On obtient alors

$$\text{Press RMCE } = \sqrt{\frac{\text{Press}}{W - p^*}}.$$

Le Press RMCE peut alors être comparé au RMCE. Une différence importante entre les deux indique que le modèle est sensible à la présence ou absence de certaines observations dans le modèle.

- **Q²** : cette statistique aussi connue comme le R² de validation croisée, n'est affichée que si l'option Press est activée. Elle est définie par :

$$Q^2 = 1 - \frac{\text{Press}}{\sum_{i=1}^{n}(y_i - \overline{y})^2}.$$

Le Q² indique la proportion de la variance totale expliquée par les variables explicatives lorsque les prédictions pour chaque observation sont calculées lorsque l'observation en question n'est pas dans l'échantillon servant à l'estimation des paramètres. Une différence importante entre le Q² et le R² indique que le modèle est sensible à l'ajout ou au retrait de certaines observations dans l'échantillon d'estimation.

Le **tableau d'analyse de la variance** permet d'évaluer le pouvoir explicatif des variables explicatives. Dans le cas où la constante du modèle n'est pas fixée à une valeur donnée, le pouvoir explicatif est évalué en comparant l'ajustement (au sens des moindres carrés) du modèle final avec l'ajustement du modèle rudimentaire composé d'une constante égale à la moyenne de la variable dépendante. Dans le cas où la constante du modèle est fixée, la comparaison est faite par rapport au modèle pour lequel la variable dépendante serait égale à la constante fixée.

Si l'option Type I/III SS (SS : Sum of Squares) est activée, les tableaux suivants sont affichés.

Le tableau des **Type I SS** permet de visualiser l'influence de l'ajout progressif des variables explicatives sur l'ajustement du modèle, au sens de la somme des carrés des erreurs (SCE), de la moyenne des carrés des erreurs (MCE), du F de Fisher, ou de la probabilité associée au F de Fisher. Plus la probabilité est faible, plus la contribution de la variable au modèle est importante, toutes les autres variables étant déjà dans le modèle. Remarques : l'ordre de sélection des variables dans le modèle influe sur les valeurs obtenues ; la somme des sommes des carrés de ce tableau est égal à la somme des carrés du modèle.

Le tableau des **Type II SS** permet de visualiser l'influence du retrait d'une variable explicative sur l'ajustement du modèle, toutes les autres variables étant conservées, au sens de la somme des carrés des erreurs (SCE), de la moyenne des carrés des erreurs (MCE), du F de Fisher, ou de la probabilité associée au F de Fisher. Plus la probabilité est faible, plus la contribution de la variable au modèle est importante. Remarque : dans le cas des ANOVAs déséquilibrées, l'utilisation des Type III est recommandée mais XLSTAT affiche les Type II pour les utilisateurs avancés qui voudraient disposer des Type II.

Le tableau des **Type III SS** permet de visualiser l'influence du retrait d'une variable explicative sur l'ajustement du modèle, toutes les autres variables étant conservées, au sens de la somme des carrés des erreurs (SCE), de la moyenne des carrés des erreurs (MCE), du F de Fisher, ou

de la probabilité associée au F de Fisher. Plus la probabilité est faible, plus la contribution de la variable au modèle est importante, toutes les autres variables étant déjà dans le modèle. Remarque : contrairement au cas des Type I SS, l'ordre de sélection des variables dans le modèle n'influe pas sur les valeurs obtenues, et contrairement aux Type II SS, les valeurs ne dépendent pas des effectifs des cellules (par cellule on entend une combinaison de modalités des différents facteurs), ce qui fait des Type III le test recommandé pour évaluer la contribution d'une variable.

L'**équation du modèle** est ensuite affichée pour faciliter la lecture ou la réutilisation du modèle.

Le tableau des **coefficients normalisés** (aussi appelés coefficients bêta) permet de comparer le poids relatif des variables. Plus la valeur absolue d'un coefficient est élevée, plus le poids de la variable correspondante est important. Lorsque l'intervalle de confiance autour des coefficients normalisés comprend la valeur 0 (cela est facilement visible sur le graphique des coefficients normalisés), le poids d'une variable dans le modèle n'est pas significatif.

Dans le tableau des **prédictions et résidus** sont donnés pour chaque observation, son poids, la valeur de la variable explicative qualitative s'il n'y en a qu'une, la valeur observée de la variable dépendante, la prédiction du modèle, les résidus, les résidus studentisés, les intervalles de confiance, ainsi que la prédiction ajustée si l'option correspondante a été activée dans la boîte de dialogue. Deux types d'intervalles de confiance sont affichés : un intervalle de confiance autour de la moyenne (correspondant au cas où l'on ferait la prédiction pour un nombre infini d'observations avec un ensemble de valeurs données des variables explicatives) et un intervalle autour de la prédiction ponctuelle (correspondant au cas d'une prédiction isolée pour des valeurs données des variables explicatives). Le second intervalle est toujours plus grand que le premier, la variabilité étant plus importante. Si des données de validation ont été sélectionnées, elles sont affichées en fin de tableau.

Les **graphiques** qui suivent permettent de visualiser les résultats mentionnés ci-dessus. S'il n'y a qu'une seule variable explicative quantitative dans le modèle et un seul facteur, le premier graphique affiché permet de visualiser les valeurs observées et les droites de régression pour chacune des modalités du facteur. Le second graphique permet quant à lui de visualiser les résidus normalisés en fonction de la variable explicative. En principe, les résidus doivent être distribués de manière aléatoire autour de l'axe des abscisses. L'observation d'une tendance ou d'une forme révèlerait un problème au niveau du modèle.

Les **trois graphiques** affichés ensuite permettent de visualiser respectivement l'évolution des résidus normalisés en fonction de la variable dépendante, la distance entre les prédictions et les observations (pour un modèle idéal, les points seraient tous sur la bissectrice), et les résidus normalisés sur la forme d'un diagramme en bâtons. Ce dernier graphique permet de rapidement voir si un nombre anormal de données sort de l'intervalle]-2, 2[sachant que ce dernier, sous hypothèse de normalité, doit contenir environ 95% des données.

Dans le tableau des **diagnostics d'influence** sont affichés pour chaque observation, son poids, le résidu, le résidu normalisé (division par la RMCE), le résidu studentisé, le résidu supprimé (Deleted), le résidu supprimé studentisé, le leverage centré, la distance de Mahalanobis, le D de Cook, le CovRatio, le DFFits, le DFFits standardisé, les DFBeta (un par coefficient du modèle) et les DFBeta standardisés.

Quatre graphiques sont ensuite affichés pour mettre en évidence les observations dont l'influence nécessite une attention particulière.

Si vous avez sélectionné des données à utiliser pour calculer des **prédictions sur de nouvelles observations**, le tableau correspondant est ensuite affiché.

Si les tests pour la vérification des hypothèses de **normalité** et d'**homoscédasticité** ont été demandés, les résultats sont ensuite affichés.

Si des tests de **comparaisons multiples** ont été demandés, les résultats correspondants sont ensuite affichés.

Lorsqu'une ANOVA à un facteur a été appliquée et que l'option correspondante a été activée, les résultats des tests de Welch et de Brown-Forsythe sont affichés. On peut retrouver les statistiques associées, les degrés de libertés ainsi que les p-valeurs.

Si plusieurs variables dépendantes ont été sélectionnées et si l'option de comparaisons multiples a été activée, un tableau indiquant les moyennes de chaque modalité de chaque facteur et pour tous les Y est affiché. Les cellules du tableau sont colorées en utilisant une échelle spectrale allant du bleu au rouge. Si il ya plus de 10 catégories, seules les 5 moyennes les plus faibles et 5 plus élevées sont colorées. Un graphique permet de visualiser les mêmes résultats.

Un deuxième graphique permet de visualiser les moyennes estimées accompagnées des lettres de regroupement issues des tableaux de comparaisons multiples. Ce deuxième graphique n'affiche les résiultats que pour les variables dépendantes significatives pour lesquelles le test F est significatif.

Si des tests de comparaison multiples ont été demandés, les résultats correspondant sont ensuite affichés.

Exemple

Un exemple d'ANCOVA est disponible sur le Centre d'aide XLSTAT à l'adresse suivante :

http://www.xlstat.com/demo-ancof.htm

Bibliographie

Akaike H. (1973). Information theory and the extension of the maximum likelihood principle. In: Second International Symposium on Information Theory. (Eds: V.N. Petrov and F. Csaki). Academiai Kiadó, Budapest. 267-281.

Amemiya T. (1980). Selection of regressors. *International Economic Review*, **21**, 331-354.

Dempster A.P. (1969). Elements of Continuous Multivariate Analysis. Addison-Wesley, Reading.

Hsu J.C. (1996). Multiple Comparisons: Theory and Methods. CRC Press, Boca Raton.

Jobson J. D. (1999). Applied Multivariate Data Analysis: Volume 1: Regression and Experimental Design. Springer Verlag, New York.

Lea P., Naes T. and Robotten M. (1997). Analysis of Variance for Sensory Data. John Wiley and Sons, London.

Mallows C.L. (1973). Some comments on Cp. *Technometrics*, **15**, 661-675.

Rogers W. H. (1993). Regression standard errors in clustered samples. _Stata Technical _Bulletin, **13**, 19–23.

Sahai H. and Ageel M.I. (2000). The Analysis of Variance. Birkhaüser, Boston.

Tomassone R., Audrain S., Lesquoy de Turckheim E. and Miller C. (1992). La Régression, Nouveaux Regards sur une Ancienne Méthode Statistique. INRA et MASSON, Paris.

Velleman P.F. and R.E. Welsch (1981). Efficient computing of regression diagnostics. *The American Statistician*, **35**, 234-242.

Welch B. L. (1951). On the comparison of several mean values: An alternative approach. *Biometrika*, **38**, 330-336.

Welsch R.E. and Kuh E. (1977). Linear Regression Diagnostics. *Sloan School of Management Working Paper*,923-977, M.I.T., Cambridge, Mass.

White H. (1980). A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity. *Econometrica*, **48(4)**, 817-838.

Zeileis A. (2006). Object-Oriented Computation of Sandwich Estimators. *Journal of Statistical Software*,**16(9)**, 1-16.

ANOVA à mesures répétées

Utilisez cet outil pour appliquer un modèle d'analyse de la variance lorsque la variable dépendante est observée plusieurs fois.

Les options permettent de choisir les contraintes sur le modèle et de prendre en compte les interactions entre les facteurs. XLSTAT propose différents types de traitement des mesures répétées. La méthode classique basée sur les moindres carrés (comme l'analyse de la variance classique) et la méthode alternative basée sur les modèles mixtes et le maximum de vraisemblance.

Dans le cadre de cette aide, nous développerons le cas basé sur les moindres carrés (LS). Pour une aide sur le cas basé sur le maximum de vraisemblance, voir le chapitre suivant sur les modèles mixtes.

Dans cette section :

Description

L'analyse de variance à mesures répétées utilise le même cadre conceptuel que l'analyse de la variance classique. Les variables explicatives sont qualitatives et la variable dépendante est quantitative et est mesurée plusieurs fois. Chacune de ces mesures est appelée répétition. Dans le cadre de l'ANOVA à mesures répétées, les variables explicatives sont souvent appelées facteurs.

Si p est le nombre de facteurs, le modèle de l'ANOVA au temps t s'écrit de la manière suivante :

$$y_i^t = \beta_0 + \sum_{j=1}^{p} \beta_{k(i,j),j} + \epsilon_i \quad (1)$$

où y_i^t est la valeur observée pour la variable dépendante pour l'observation i au temps t, $k(i,j)$ est l'indice correspondant à la modalité du facteur j pour l'observation i, et ϵ_i est l'erreur du modèle.

Les hypothèses utilisées en ANOVA à mesures répétées sont identiques à celles de la régression linéaire : les erreurs ϵ_i suivent une même loi normale $N(0, s)$ et sont indépendantes.

Néanmoins, certaines hypothèses supplémentaires sont nécessaires. Comme les différentes répétitions sont effectuées sur les mêmes individus, on ne pourra pas supposer que y^{t_1} et y^{t_2} sont indépendantes, on supposera néanmoins que la matrice de covariance entre les y est sphérique. Cette hypothèse peut être relâchée lorsqu'on utilise les modèles mixtes.

Le principe de l'ANOVA sur mesures répétées est simple, on va effectuer T ANOVA classiques sur chaque répétition $t_1, \ldots, t_T$ et ensuite on va tester la sphéricité de la matrice de covariance entre les répétitions en utilisant le test de Mauchly et les epsilons de Greenhouse-Geisser et Huynt-Feldt. Si l'hypothèse de sphéricité est vérifiée, on pourra analyser les tests sur les effets intra- et inter-sujets.

Interactions

On désigne par interaction un facteur artificiel (non mesuré) reflétant l'interaction entre au moins deux facteurs mesurés. Par exemple, si on applique un traitement à une plante, et que les essais sont réalisés sous deux intensités lumineuses différentes, on pourra inclure dans le modèle un facteur d'interaction traitement*lumière qui permettra d'identifier une éventuelle interaction entre les deux facteurs. S'il y a une interaction entre les deux facteurs, on observera sur les plantes un effet significativement plus important lorsque la lumière est forte et que le traitement est de type 2, alors que l'effet est moyen pour les couples (lumière faible, traitement 2) et (lumière forte, traitement 1).

Pour faire un parallèle avec la régression linéaire, les interactions sont équivalentes à des produits entre les valeurs explicatives continues, bien qu'ici l'obtention des interactions nécessite plus qu'une simple multiplication entre deux variables. Néanmoins la notation utilisée pour représenter l'interaction entre le facteur A et le facteur B est A*B.

XLSTAT permet de facilement définir les interactions à prendre en compte dans le modèle.

Facteurs imbriqués

Lorsqu'on ne peut pas croiser toutes les modalités de deux facteurs, alors on peut utiliser des facteurs imbriqués ou hiérarchiques. Par exemple, si l'on cherche à analyser le lien entre certaines caractéristiques d'un produit en sortie d'une chaîne de fabrication et les opérateurs et les machines impliqués, et si les opérateurs travaillent sur une machine donnée avec par exemple quatre opérateurs en rotation, alors chaque opérateur n'est pas croisé avec chaque machine, mais est associé à une seule machine. On a alors un effet imbriqué (l'effet opérateur est imbriqué dans l'effet machine).

XLSTAT permet d'identifier automatiquement les facteurs imbriqués. Par ailleurs, il est possible d'inclure dans le modèle un facteur imbriqué.

Contraintes

Au cours des calculs, chaque facteur est décomposé en une sous-matrice contenant autant de colonnes qu'il y a de modalités dans le facteur. Typiquement, il s'agit d'un tableau disjonctif complet. Cette décomposition pose néanmoins un problème : s'il y a g modalités, le rang de cette sous- matrice n'est pas g mais $g - 1$. Cela entraîne la nécessité de supprimer l'une des colonnes de la sous-matrice, et éventuellement de transformer les autres colonnes. Plusieurs stratégies sont possibles en fonction de l'interprétation que l'on veut ensuite faire :

1) **a1=0** : le paramètre correspondant à la première modalité est nul. Ce choix permet d'imposer que l'effet de la première modalité correspond à un standard. Dans ce cas, la constante du modèle est égale à la moyenne de la variable dépendante pour le groupe 1.

2) **an=0** : le paramètre correspondant à la dernière modalité est nul. Ce choix permet d'imposer que l'effet de la dernière modalité correspond à un standard. Dans ce cas, la constante du modèle est égale à la moyenne de la variable dépendante pour le groupe g.

Remarque : si le choix de la contrainte influence la valeur des paramètres, il n'en a aucun sur les valeurs prédites et sur les différentes statistiques d'ajustement.

Tests de comparaisons multiples

L'une des applications principales de l'ANOVA sont les tests de comparaisons multiples dont le but est de vérifier si les paramètres correspondant aux différentes modalités d'un facteur sont significativement différents ou non. Par exemple, dans le cas où quatre traitements sont appliqués à des plantes, on veut savoir non seulement si les traitements ont un effet significatif, mais aussi si les traitements ont un effet différent.

De nombreux tests ont été proposés pour comparer les moyennes des modalités. La majorité de ces tests s'appuie aussi sur l'hypothèse de normalité. XLSTAT propose les principaux tests parmi lesquels :

Test de Tukey (HSD) : ce test est le plus utilisé (HSD : *honestly significant difference*).

Test de Fisher (LSD) : c'est un test de Student qui permet de tester l'hypothèse nulle que toute les moyennes pour les différentes modalités sont égales (LSD : *least significant difference*).

Test du t* de Bonferroni : dérivé du test de Student, il est un peu plus performant car il prend en compte le fait que plusieurs comparaisons sont effectuées simultanément. En conséquence le niveau de signification du test est modifié suivant la formule suivante :

$$\alpha' = \frac{\alpha}{g(g-1)/2}$$

où g est le nombre de modalités du facteur dont les modalités sont comparées.

Test de Dunn-Sidak : dérivé du test de Bonferroni, il est plus performant dans certaines situations.

$$\alpha' = 1 - (1-\alpha)^{2/[g(g-1)]}.$$

Les tests suivants sont plus complexes et consistent en des procédures itératives pour lesquelles les résultats dépendent du nombre de combinaisons restant à tester.

Test de Newman-Keuls (SNK) : dérivé du test de Student (SNK : Student Newman-Keuls), il est très souvent utilisé bien que pas très performant.

Test de Duncan : ce test est peu utilisé.

Test de REGWQ : ce test est la procédure itérative la plus performante dans une majorité de situations (REGWQ : Ryan-Einot-Gabriel-Welsch).

La procédure de **Benjamini-Hochberg** permet de contrôler le taux de faux positifs (False Discovery Rate ou FDR). Cette procédure de pénalisation des p-values est peu conservative.

Le test de **Games-Howell** (GH) peut être utilisé dans les ANOVAs à un facteur lorsque les variances ne sont pas d'homogènes. Il peut être utilisé avec des tailles d'échantillon inégales, mais il est recommandé de l'utiliser quand le plus petit échantillon a 5 éléments ou plus, sinon il est trop libéral (au sens qu'il a tendance à rejeter trop facilement l'hypothèse H0). Le test de **T2 du Tamhane** est plus conservateur, mais pas aussi puissant que le test GH.

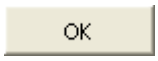
Tous les tests ci-dessus permettent de comparer toutes les paires de modalités et appartiennent à la famille des tests MCA (*Multiple Comparisons of All*, ou *All-Pairwise Comparisons*).

D'autres tests permettent de comparer toutes les catégories à une modalité témoin. Ces tests sont appelés tests MCB (*Multiple Comparisons with the Best*, *Comparisons with a control*). XLSTAT propose le test de Dunnett qui est le plus utilisé. On distingue trois tests de Dunnett :

- **Test bilatéral** : l'hypothèse nulle suppose l'égalité entre la modalité testée et la modalité témoin. L'hypothèse alternative suppose que les moyennes des deux modalités sont différentes.
- **Test unilatéral à gauche** : l'hypothèse nulle suppose l'égalité entre la modalité testée et la modalité témoin. L'hypothèse alternative suppose que la moyenne de la modalité témoin est supérieure à la moyenne de la modalité testée.
- **Test unilatéral à droite** : l'hypothèse nulle suppose l'égalité entre la modalité testée et la modalité témoin. L'hypothèse alternative suppose que la moyenne de la modalité témoin est inférieure à la moyenne de la modalité testée.

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.

Onglet **Général** :

Y / Variables dépendantes :

Quantitatives : sélectionnez la variable réponse que vous souhaitez modéliser. Si des en-têtes de colonnes ont été sélectionnés, veuillez vérifier que l'option « Libellés des variables » est activée.

Une colonne pour toutes les répétitions : activez cette option si votre variable dépendante est répartie sur une seule colonne. Dans ce cas, vous devrez sélectionner comme variables explicatives une variable donnant le libellé de la répétition et une seconde donnant le sujet traité. Voir l'aide sur les modèles mixtes pour plus de détails sur ce format.

Une colonne par répétition : activez cette option si votre variable dépendante est répartie sur T colonnes et que chaque colonne représente une répétition. Dans ce cas, lorsque vous devrez sélectionner les facteurs, un facteur répétition et un facteur sujet apparaitront.

X / Variables explicatives :

Qualitatives : sélectionnez la ou les variables explicatives qualitatives (les facteurs) sur la feuille Excel. Les données sélectionnées peuvent être de tout type, mais les données numériques sont automatiquement considérées comme nominales. Si le libellé des variables a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

Quantitatives : sélectionnez la ou les variables explicatives quantitatives sur la feuille Excel. Si le libellé des variables a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée. Lorsqu'aucune variable qualitative n'est sélectionnée, on est alors dans le cadre d'une régression linéaire à mesures répétées. Si des données qualitatives et quantitatives sont sélectionnées, on est alors dans le cas d'une ANCOVA à mesures répétées. Si finalement, on ne sélectionne aucune variable explicative, on se trouve dans le cas d'une ANOVA à mesures répétées à un facteur.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des variables : activez cette option si la première ligne des données sélectionnées (variables dépendantes et explicatives, libellés des observations, poids) contient un libellé.

Poids des observations : activez cette option si vous voulez pondérer les observations. Si vous n'activez pas cette option, les poids seront tous considérés comme valant 1. Les poids doivent être impérativement supérieurs ou égaux à 0. Un poids de 2 est équivalent à répéter

deux fois la même observation. Si un en-tête de colonne a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

Poids dans la régression : activez cette option si vous voulez effectuer une régression par les moindres carrés pondérés. Si vous n'activez pas cette option, les poids seront tous considérés comme valant 1. Les poids doivent être impérativement supérieurs ou égaux à 0. Si un en-tête de colonne a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

Onglet **Options** :

Constante fixée : activez cette option pour fixer la constante du modèle de régression à une valeur que vous devez ensuite saisir (0 par défaut).

Tolérance : activez cette option pour permettre à l'algorithme de calcul de la régression OLS ne pas prendre en compte les variables qui seraient soit constantes soit trop corrélées avec d'autres variables déjà utilisées dans le modèle (0.0001 par défaut).

Interactions / Niveau : activez cette option pour inclure des interactions dans le modèle puis entrez le niveau maximum d'interaction (valeur comprise entre 1 et 4).

Effets imbriqués : activez cette option pour inclure un effet imbriqué dans le modèle.

Intervalle de confiance (%) : entrez l'étendue en pourcentage de l'intervalle de confiance à utiliser pour les différents tests, et pour le calcul des intervalles de confiance autour des paramètres et des prédictions. Valeur par défaut : 95.

Méthode d'estimation : trois méthodes sont disponibles. La première est la méthode classique basée sur les moindres carrés et notée LS. Les deux autres méthodes sont basées sur le maximum de vraisemblance et sont développées dans le cadre de l'aide sur les modèles mixtes (les sorties sont alors complètements différentes).

Contraintes : des détails sur les différentes options sont disponibles dans la section description.

- **a1 = 0** : choisissez cette option pour que le paramètre de la première modalité de chaque facteur soit fixé à 0.
- **an = 0** : choisissez cette option pour que le paramètre de la dernière modalité de chaque facteur soit fixé à 0.

Onglet **Données manquantes** :

Supprimer les observations : activez cette option pour supprimer les observations comportant des données manquantes.

Estimer les données manquantes : activez cette option pour estimer les données manquantes avant le début des calculs.

- **Moyenne ou mode** : activez cette option pour estimer les données manquantes en utilisant la moyenne (variables quantitatives) ou le mode (variables qualitatives) pour les variables correspondantes.
- **Plus proche voisin** : activez cette option pour estimer les données manquantes d'une observation en recherchant le plus proche voisin de l'observation.

Onglet **Sorties** :

Statistiques descriptives : activez cette option pour afficher les statistiques descriptives pour les variables sélectionnées.

Analyse de la variance : activez cette option pour afficher le tableau d'analyse de la variance pour chaque analyse de la variance associé à la répétition t.

Type I/III SS : activez cette option pour afficher les tableaux Type I SS et Type III SS permettant de mesurer la contribution des différentes variables explicatives au modèle (SS correspond à *Sum of Squares*) pour chaque analyse de la variance associé à la répétition t.

Coefficients normalisés : activez cette option pour afficher les paramètres normalisés du modèle (coefficients bêta).

Tests des effets intra-sujets : activez cette option pour afficher les tests sur les effets intra-sujet.

Tests des effets inter-sujets : activez cette option pour afficher les tests sur les effets inter-sujet.

Test de sphéricité de Mauchly : Activez cette option pour afficher le test de sphéricité de Mauchly.

Prédictions et résidus : activez cette option pour afficher les prédictions et les résidus pour l'ensemble des observations.

Comparaisons multiples :

Des informations sur les tests de comparaisons multiples sont disponibles dans la section description.

Appliquer à tous les facteurs : activez cette option pour calculer les tests sélectionnés pour tous les facteurs.

Utiliser les moyennes estimées : activez cette option pour calculer les moyennes en utilisant le modèle. Si cette option n'est pas activée, les moyennes sont estimées à partir des données.

Trier en ordre croissant : activez cette option pour trier les modalités comparées en ordre croissant, le critère de tri étant leur moyenne respective. Si cette option n'est pas activée, le tri est décroissant.

Comparaison par paires : activez cette option puis choisissez les méthodes de comparaison.

Comparaison à un témoin : activez cette option puis choisissez le type de test de Dunnett que vous voulez effectuer.

Onglet **Graphiques** :

Graphiques de régression : activez cette option pour afficher les graphiques de régression :

- **Prédictions et résidus** : activez cette option pour afficher les graphiques suivants :

  (1) Droite de régression : ce graphique n'est affiché que s'il n'y a qu'une seule variable explicative, et que cette variable est quantitative.

  (2) Variable explicative versus résidus normalisés : ce graphique n'est affiché que s'il n'y a qu'une seule variable explicative, et que cette variable est quantitative.

  (3) Variable dépendante versus résidus normalisés.

  (4) Prédictions pour la variable dépendante versus variable dépendante.

  (5) Graphique en bâtons des résidus normalisés.

Graphiques des moyennes : activez cette option pour afficher les graphiques permettant de visualiser les moyennes pour les différentes modalités des différents facteurs.

Boîte de dialogue facteurs et interactions

Une fois la première boîte de dialogue fermée, une seconde boîte apparaît. Celle-ci est intitulée facteurs et interactions.

Il faut sélectionner les facteurs et interactions (effets fixes), un facteur répété et un facteur sujet. Si le format « une colonne par répétition » a été sélectionné, alors deux nouveaux facteurs (appelés respectivement répétition et sujet) apparaissent et doivent être sélectionnés en tant que facteur répété et facteur sujet.

Résultats

Statistiques descriptives : les tableaux de statistiques descriptives présentent pour toutes les variables sélectionnées des statistiques simples. Pour les variables dépendantes (en bleu) et les variables explicatives quantitatives, sont affichés le nombre d'observations, le nombre de données manquantes, le nombre de données non manquantes, la moyenne, et l'écart-type (non biaisé). Pour les variables explicatives qualitatives sont affichés le nom des différentes modalités ainsi que leur fréquence respective.

Ensuite pour chaque répétition, nous avons :

Coefficients d'ajustement : dans ce tableau sont affichées les statistiques relatives à l'ajustement du modèle de régression :

- **Observations** : le nombre d'observations prises en compte dans les calculs. Dans les formules présentées ci-dessous n désigne le nombre d'observations.

- **Somme des poids** : la somme des poids des observations prises en compte dans les calculs. Dans les formules présentées ci-dessous W désigne la somme des poids.

- **DDL** : le nombre de degrés de liberté pour le modèle retenu (correspondant à la partie erreurs).

- **R²** : le coefficient de détermination du modèle. Ce coefficient, dont la valeur est comprise entre 0 et 1, n'est affiché que si la constante du modèle n'est pas fixée par l'utilisateur. Sa valeur est définie par $$ R^ { 2} = \frac { \sum _ { i = 1} ^ { n } w _ { i } (\hat {y} _ { i } - \bar { y } _ { i }) ^ { 2} } { \sum _ { i = 1} ^ { n } w _ { i } (y _ { i } - \bar { y } _ { i }) ^ { 2} } \text { with } \bar {y}= \frac {1}{W} \sum _ { i = 1} ^ { n } w _ { i } { y _ i } $$ Le R² s'interprète comme la proportion de la variabilité de la variable dépendante expliquée par le modèle. Plus le R² est proche de 1, meilleur est le modèle. L'inconvénient du R² est qu'il ne prend pas en compte le nombre de variables utilisées pour ajuster le modèle.

- **R²ajusté** : le coefficient de détermination ajusté du modèle. Le R² ajusté peut être négatif si le R² est voisin de zéro. Ce coefficient n'est affiché que si la constante du modèle n'est pas fixée par l'utilisateur. Sa valeur est définie par $$ \hat { R } ^ { 2} = 1- (1- R ^ { 2}) \frac { W - 1} { W - p - 1} $$ Le R² ajusté est une correction du R² qui permet de prendre en compte le nombre de variables utilisées dans le modèle.

- **MCE** : la moyenne des carrés des erreurs (MCE) est définie par :

$$MCE = \frac{1}{W - p^*} \sum_{i=1}^{n} w_i (y_i - \hat{y}_i)^2.$$

- **RMCE** : la racine de la moyenne des carrés des erreurs (RMCE) est la racine carrée de la MCE.

- **MAPE** : la *Mean Absolute Percentage Error* est calculée comme suit :

$$MAPE = \frac{100}{W} \sum_{i=1}^{n} w_i \left| \frac{y_i - \hat{y}_i}{y_i} \right|.$$

- **DW** : le coefficient de Durbin-Watson est défini par $$ D W = \frac { \sum _ { i = 2} ^ { x } [(y _ { i } - \hat { y } _ { i }) - (y _ { i - 1} - \hat { y } _ { i - 1})] ^ { 2} } { \sum _ { i = 1} ^ { x } w _ { i } (y _ { i } - \hat { y } _ { i }) ^ { 2} }\cdot $$ Ce coefficient correspond au coefficient d'autocorrélation d'ordre 1 et permet de vérifier que les résidus du modèle ne sont pas autocorrélés, sachant que l'indépendance des résidus est l'une des hypothèses de base de la régression linéaire. L'utilisateur pourra se référer à une table des coefficients de Durbin-Watson pour vérifier si l'hypothèse d'indépendance des résidus est acceptable.

- **Cp** : le coefficient Cp de Mallows est défini par

  $$ C p = \frac { S C E } { \hat { \sigma } } + 2p ^ { * } - W , $$ où SCE est la somme du carré des erreurs pour le modèle avec p variables explicatives, et où $\hat{\sigma}$ correspond à l'estimateur de la variance des résidus pour le modèle comprenant toutes les variables explicatives. Plus le coefficient Cp est proche de p* moins le modèle est biaisé.

- **AIC** : le critère d'information d'Akaike (Akaike's Information Criterion) est défini par

$$ A I C = W \operatorname{ln} (\frac { S C E } { W }) + 2p ^ { * }\cdot $$ Ce critère proposé par Akaike (1973) dérive de la théorie de l'information, et s'appuie sur la mesure de Kullback et Leibler (1951). C'est un critère de sélection de modèles qui pénalise les modèles pour lesquels l'ajout de nouvelles variables explicatives n'apporte pas suffisamment d'information au modèle, l'information étant mesurée au travers de la SCE. On cherche à minimiser le critère AIC.

- **SBC** : le critère bayésien de Schwarz (Schwarz's Bayesian Criterion) est défini par

  $$ S B C = W \operatorname{ln} (\frac { S C E } { W }) + \operatorname{ln} (W) p ^ { * }\cdot $$ Ce critère proposé par Schwarz (1978) est proche du critère AIC, et comme ce dernier on cherche à le minimiser.

- **PC** : le critère de prédiction d'Amemiya (Amemiya's Prediction Criterion) est défini par

  $$ P C = \frac { (1- R ^ { 2}) (W + p ^ { * }) } { W - p ^ { * } }\cdot $$ Ce critère proposé par Amemiya (1980) permet comme le R² ajusté de tenir compte de la parcimonie du modèle.

Le **tableau d'analyse de la variance** permet d'évaluer le pouvoir explicatif des variables explicatives. Dans le cas où la constante du modèle n'est pas fixée à une valeur donnée, le pouvoir explicatif est évalué en comparant l'ajustement (au sens des moindres carrés) du modèle final avec l'ajustement du modèle rudimentaire composé d'une constante égale à la moyenne de la variable dépendante. Dans le cas où la constante du modèle est fixée, la comparaison est faite par rapport au modèle pour lequel la variable dépendante serait égale à la constante fixée.

Si l'option Type I/III SS (SS : Sum of Squares) est activée, les tableaux suivants sont affichés.

Le tableau des **Type I SS** permet de visualiser l'influence de l'ajout progressif des variables explicatives sur l'ajustement du modèle, au sens de la somme des carrés des erreurs (SCE), de la moyenne des carrés des erreurs (MCE), du F de Fisher, ou de la probabilité associée au F de Fisher. Plus la probabilité est faible, plus la contribution de la variable au modèle est importante, toutes les autres variables étant déjà dans le modèle. Remarque : l'ordre de sélection des variables dans le modèle influe sur les valeurs obtenues.

Le tableau des **Type III SS** permet de visualiser l'influence du retrait d'une variable explicative sur l'ajustement du modèle, toutes les autres variables étant conservées, au sens de la somme des carrés des erreurs (SCE), de la moyenne des carrés des erreurs (MCE), du F de Fisher, ou de la probabilité associée au F de Fisher. Plus la probabilité est faible, plus la contribution de la variable au modèle est importante, toutes les autres variables étant déjà dans le modèle. Remarque : contrairement au cas des Type I SS, l'ordre de sélection des variables dans le modèle n'influe pas sur les valeurs obtenues, et contrairement aux Type II SS, les valeurs ne dépendent pas des effectifs des cellules (par cellule on entend une combinaison de modalités des différents facteurs), ce qui fait des Type III le test recommandé pour évaluer la contribution d'une variable.

Le tableau **paramètres du modèle** affiche l'estimation des paramètres, l'erreur standard correspondante, le t de Student, la probabilité correspondante, ainsi que l'intervalle de confiance.

Le tableau des **coefficients normalisés** (aussi appelés coefficients bêta) permet de comparer le poids relatif des variables. Plus la valeur absolue d'un coefficient est élevée, plus le poids de la variable correspondante est important. Lorsque l'intervalle de confiance autour des coefficients normalisés comprend la valeur 0 (cela est facilement visible sur le graphique des coefficients normalisés), le poids d'une variable dans le modèle n'est pas significatif.

Dans le tableau des **prédictions et résidus** sont donnés pour chaque observation, son poids, la valeur de la variable explicative qualitative s'il n'y en a qu'une, la valeur observée de la variable dépendante, la prédiction du modèle, les résidus, les intervalles de confiance. Deux types d'intervalles de confiance sont affichés : un intervalle de confiance autour de la moyenne (correspondant au cas où l'on ferait la prédiction pour un nombre infini d'observations avec un ensemble de valeurs données des variables explicatives) et un intervalle autour de la prédiction ponctuelle (correspondant au cas d'une prédiction isolée pour des valeurs données des variables explicatives). Le second intervalle est toujours plus grand que le premier, les aléas étant plus importants. Si des données de validation ont été sélectionnées, elles sont affichées en fin de tableau.

Les **graphiques** qui suivent permettent de visualiser les résultats mentionnés ci-dessus. S'il n'y a qu'une seule variable explicative dans le modèle, le premier graphique affiché permet de visualiser les valeurs observées, la droite de régression et les deux types d'intervalles de confiance autour des prévisions. Le second graphique permet quant à lui de visualiser les résidus normalisés en fonction de la variable explicative. En principe, les résidus doivent être distribués de manière aléatoire autour de l'axe des abscisses. L'observation d'une tendance ou d'une forme révèlerait un problème au niveau du modèle.

Les trois graphiques affichés ensuite permettent de visualiser respectivement l'évolution des résidus normalisés en fonction de la variable dépendante, la distance entre les prédictions et les observations (pour un modèle idéal, les points seraient tous sur la bissectrice), et les résidus normalisés sur la forme d'un diagramme en bâtons. Ce dernier graphique permet de rapidement voir si un nombre anormal de données sort de l'intervalle]-2, 2[sachant que ce dernier, sous hypothèse de normalité, doit contenir environ 95% des données.

Si des tests de comparaison multiples ont été demandés, les résultats correspondant sont ensuite affichés.

Finalement, les tableaux associés aux mesures répétées sont présentés :

Le test de sphéricité de Mauchly permet de valider l'hypothèse de sphéricité. La puissance de ce test est assez faible et il pourra poser des problèmes pour de petits échantillons. Dans ce tableau, on trouve aussi l'epsilon de Greenhouse-Geisser et l'epsilon de Huynt-Feldt. Ces deux statistiques permettent aussi de tester l'hypothèse de sphéricité. Plus elles sont proches de 1, plus on se rapprochera d'une matrice de covariance sphérique.

Test sur les effets intra-sujets : Ce tableau nous permet de voir quels facteurs ont un effet qui évolue d'une répétition à une autre.

Test sur les effets inter-sujets : Ce tableau nous permet de voir quels facteurs ont un effet qui diffère d'un sujet à un autre et non d'une répétition à une autre.

Exemple

Un exemple d'ANOVA avec mesures répétées est disponible sur le Centre d'aide XLSTAT à l'adresse suivante :

http://www.xlstat.com/demo-anorep2f.htm

Bibliographie

Akaike H. (1973). Information theory and the extension of the maximum likelihood principle. In: Second International Symposium on Information Theory. (Eds: V.N. Petrov and F. Csaki). Academiai Kiadó, Budapest. 267-281.

Dempster A.P. (1969). Elements of Continuous Multivariate Analysis. Addison-Wesley, Reading.

Girden E.R. (1992). ANOVA Repeated Measures. Sage University Paper.

Greenhouse S.W., Geisser S. (1959). On methods in the analysis of profile data. *Psychometrika*. 24, 95-112.

Hsu J.C. (1996). Multiple Comparisons: Theory and Methods. CRC Press, Boca Raton.

Huynt H., Feldt L.S. (1976). Estimation of the Box correction for degrees of freedom from sample data i, randomized block and split-plot designs. *Journal of Educational Statistics*. 1, 69-82.

Jobson J. D. (1999). Applied Multivariate Data Analysis: Volume 1: Regression and Experimental Design. Springer Verlag, New York.

Lea P., Naes T. and Robotten M. (1997). Analysis of Variance for Sensory Data. John Wiley and Sons, London.

Mallows C.L. (1973). Some comments on Cp. *Technometrics*, **15**, 661-675.

Mauchly, J.W. (1940). Significance test for sphericity of n-variate normal population. *Annals of Mathematical Statistics*. 11, 204-209.

Sahai H. and Ageel M.I. (2000). The Analysis of Variance. Birkhaüser, Boston.

Searle, S. R., Casella, G., and McCulloch, C. E. (1992). Variance Components. John Wiley & Sons, New York.

Modèles mixtes

Utilisez cet outil pour appliquer les modèles mixtes à facteurs répétés ou à facteurs aléatoires. Ces méthodes permettent entre autres d'appliquer l'analyse de la variance sur mesures répétées ainsi que les modèles à facteurs aléatoires.

Dans cette section :

Description

Les modèles mixtes sont des modèles complexes qui ont été développés à partir du modèle linéaire. Ils permettent de prendre en compte, d'une part, la notion de mesure répétée et, d'autre part, celle de facteur aléatoire. Les variables explicatives pourront être aussi bien quantitatives que qualitatives. Dans le cadre des modèles mixtes, les variables explicatives sont souvent appelées facteurs. XLSTAT permet de faire une ANOVA sur mesures répétées en utilisant les modèles mixtes.

L'équation du modèle aura donc la forme suivante :

$$y = X\beta + Z\gamma + \epsilon \quad (1)$$

où y est la variable quantitative à expliquer, X rassemble les facteurs associés aux effets fixes (ce sont les variables classiques de la régression linéaire), β est un vecteur de coefficients associés aux effets fixes, Z est une matrice rassemblant les effets aléatoires (ce sont des variables qui ne peuvent pas être supposées fixes), γ est un vecteur de coefficients associés aux effets aléatoire et ϵ est un vecteur rassemblant les erreurs associées à chaque observation. A la différence du modèle linéaire classique, nous avons : $\gamma \sim N(0, G(\theta_G))$ et $\epsilon \sim N(0, R(\theta_R))$.

On a donc :

$$E\begin{bmatrix}\gamma \\ \epsilon\end{bmatrix} = \begin{bmatrix}0 \\ 0\end{bmatrix} \text{ et } Var\begin{bmatrix}\gamma \\ \epsilon\end{bmatrix} = \begin{bmatrix}G & 0 \\ 0 & R\end{bmatrix}$$

La variance de y est donc donnée par la formule : $Var(y) = V(\theta) = Z'GZ + R$, où θ est un vecteur de paramètres associés aux inconnues des matrices de covariance G et R. On a donc $y \sim N(X\beta, V(\theta))$.

En fonction du modèle à estimer, les matrices R et G auront des formes différentes :

- Dans le cas d'un modèle linéaire classique, on aura : $Z = 0$ et $R = \sigma^2 I_n$.
- Si on veut faire une ANOVA avec mesures répétées, on aura $Z = 0$ et $cov(\epsilon) = R(\theta)$, où R est une matrice par blocs dont la forme sera définie par l'utilisateur. Ainsi, chaque bloc rassemble les covariances entre les différentes mesures effectuées sur un sujet. Les variables explicatives seront toutes qualitatives.
- Si on veut appliquer un modèle à effet aléatoire, on aura donc une matrice Z rassemblant les effets aléatoires et $cov(\gamma) = G$, où G est une matrice dont la structure est définie par l'utilisateur.

Nous rassemblons dans le tableau suivant les différentes structures de covariance pour R et G présentes dans XLSTAT avec p taille de la matrice (les matrices sont symétriques) :

Structure de covariance	Nombre de paramètres	Formule
Variance components	Nombre d'effets aléatoire (si pas d'effet aléatoire = 1)	$\sigma_{ij} = \sigma_k^2 I_{(i=j)}$, k est l'effet associé à la ligne i
Autoregressive(1)	2	$\sigma_{ij} = \sigma^2 \rho^{\lvert i-j \rvert}$
Compound symmetry	2	$\sigma_{ij} = \sigma_1 + \sigma^2 I_{(i=j)}$
Unstructured	$p(p+1)/2$	$\sigma_{ij} = \sigma_{ij}$
Toeplitz	p	$\sigma_{ij} = \sigma_{\lvert i-j \rvert+1}$
Toeplitz(q)	Min(p,q)	$\sigma_{ij} = \sigma_{\lvert i-j \rvert+1} I_{(\lvert i-j \rvert<q)}$

L'estimation des paramètres q se fait en utilisant le principe du maximum de vraisemblance. Il existe deux méthodes d'estimation : le maximum de vraisemblance classique (ML) et le maximum de vraisemblance restreint (REML), cette dernière est celle utilisée par défaut. La fonction de vraisemblance est donc :

$$l_{REML}(G,R) = -\frac{1}{2} log \mid V \mid -\frac{1}{2} log \mid X'V^{-1}X \mid -\frac{1}{2} r'V^{-1}r - \frac{n-p}{2} log(2\pi) \quad (2)$$

où $r = y - X\hat{\beta}$. Les paramètres sont obtenus en utilisant les dérivées premières et secondes de $l_{REML}(G,R)$. Pour les détails de l'obtention de ces matrices, on peut voir Wolfinger, Tobias et Sall (1994). XLSTAT ne fait pas de profilage de la variance dans ses calculs et utilise

l'estimation de la variance obtenue par le modèle linéaire général comme valeurs initiales pour les matrices de covariance. L'utilisation d'une méthode analytique pour obtenir les paramètres θ n'est pas possible. Nous utilisons donc l'algorithme itératif de Newton-Raphson afin d'obtenir une estimation de θ. Une fois θ obtenu, les coefficients β et γ sont calculés en résolvant le système d'équations suivant :

$$\begin{bmatrix} X'\hat{R}^{-1}X & X'\hat{R}^{-1}Z \\ Z'\hat{R}^{-1}X & Z'\hat{R}^{-1}Z + \hat{G}^{-1} \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{\gamma} \end{bmatrix} = \begin{bmatrix} X'\hat{R}^{-1}y \\ Z'\hat{R}^{-1}y \end{bmatrix} \tag{3}$$

On obtient donc :

$$\begin{aligned} \hat{\beta} &= \left(X'\hat{V}^{-1}X\right)^{-} X'\hat{V}^{-1}y \\ \hat{\gamma} &= \hat{G}Z'\hat{V}^{-1}\left(y - X\hat{\beta}\right) \end{aligned} \tag{4}$$

où $()^{-}$ est l'inverse généralisée de la matrice entre parenthèses. L'interprétation du modèle se fait de la même façon que dans le cas linéaire.

Format des données

Dans le cadre des modèles mixtes, les données auront un format spécifique :

S'il n'y a pas de mesures répétées, alors on aura une colonne par variable associée à chaque effet fixe et une colonne par variable associée à chaque effet aléatoire.

Si on a des mesures répétées, l'ensemble des répétitions devra se trouver à la suite les unes des autres. On ne pourra pas avoir une colonne par répétition. On définit donc un facteur permettant d'identifier chaque répétition et un autre facteur permettant d'identifier le sujet traité lors de chaque répétition. Ainsi sur un jeu de données avec 3 répétitions et 2 sujets et pour une variable explicative X mesurée aux temps T_1 et T_2 sur les deux sujets, on aura le tableau suivant :

fact.rep	fact.suj	X
1	1	$x_1^{T_1}$
1	2	$x_1^{T_2}$
2	1	$x_1^{T_2}$
2	2	$x_2^{T_1}$
3	1	$x_1^{T_3}$
3	2	$x_2^{T_3}$

XLSTAT permet de sélectionner un facteur répété et un facteur sujet. Ces facteurs doivent être qualitatifs.

Interactions

On désigne par interaction un facteur artificiel (non mesuré) reflétant l'interaction entre au moins deux facteurs mesurés. Par exemple, si on applique un traitement à une plante, et que les essais sont réalisés sous deux intensités lumineuses différentes, on pourra inclure dans le modèle un facteur d'interaction traitement*lumière qui permettra d'identifier une éventuelle interaction entre les deux facteurs. S'il y a une interaction entre les deux facteurs, on observera sur les plantes un effet significativement plus important lorsque la lumière est forte et que le traitement est de type 2, alors que l'effet est moyen pour les couples (lumière faible, traitement 2) et (lumière forte, traitement 1).

Pour faire un parallèle avec la régression linéaire, les interactions sont équivalentes à des produits entre les valeurs explicatives continues, bien qu'ici l'obtention des interactions nécessite plus qu'une simple multiplication entre deux variables. Néanmoins la notation utilisée pour représenter l'interaction entre le facteur A et le facteur B est A*B.

XLSTAT permet de facilement définir les interactions à prendre en compte dans le modèle.

Contraintes

Au cours des calculs, chaque facteur est décomposé en une sous-matrice contenant autant de colonnes qu'il y a de modalités dans le facteur. Typiquement, il s'agit d'un tableau disjonctif complet. Cette décomposition pose néanmoins un problème : s'il y a g modalités, le rang de cette sous- matrice n'est pas g mais $g-1$. Cela entraîne la nécessité de supprimer l'une des colonnes de la sous-matrice, et éventuellement de transformer les autres colonnes. Plusieurs stratégies sont possibles en fonction de l'interprétation que l'on veut ensuite faire :

1) **a1=0** : le paramètre correspondant à la première modalité est nul. Ce choix permet d'imposer que l'effet de la première modalité correspond à un standard. Dans ce cas, la constante du modèle est égale à la moyenne de la variable dépendante pour le groupe 1.

2) **an=0** : le paramètre correspondant à la dernière modalité est nul. Ce choix permet d'imposer que l'effet de la dernière modalité correspond à un standard. Dans ce cas, la constante du modèle est égale à la moyenne de la variable dépendante pour le groupe g.

Remarque : si le choix de la contrainte influence la valeur des paramètres, il n'en a aucun sur les valeurs prédites et sur les différentes statistiques d'ajustement.

Inférence et tests

XLSTAT permet d'afficher les tests de type I, II et III sur les effets fixes. Le principe de ces tests est le même que dans le cas du modèle linéaire. Néanmoins, leur calcul diffère légèrement. L'ensemble de ces tests est basé sur la statistique F suivante :

$$F = \frac{\hat{\beta}' L' (L(X'\hat{V}^{-1}X)^{-}L')^{-}L\hat{\beta}}{r}$$

où L est une matrice construite spécialement dans le cas de chaque type d'erreur. De plus, $r = rang(L(X'\hat{V}^{-1}X)^{-}L')$. Une p-valeur peut être obtenue en utilisant la distribution de Fisher avec $Num.DDL$ et $Den.DDL$ degrés de liberté. On a $Num.DDL = rang(L)$ et l'estimation de $Den.DDL$ dépendra du modèle sélectionné. XLSTAT utilise :

- La méthode *contain* si au moins un effet aléatoire est sélectionné, on a : $Den.DDL = N - rang(XZ)$.
- La méthode *residual* si il n'y a pas d'effet aléatoire : $\text{Den}.DDL = n - \text{rang}(X)$.
- La méthode *approximation de Satterthwaite* dans le cas ou les données sont non équilibrées : $\text{Den}.DDL = \frac{2E}{E-q}$.

$$\text{avec } E = \sum_{m=1}^{q} \frac{v_m}{v_m - 2} I(v_m > 2); \quad v_m = \frac{2(D_m)^2}{g_m' A g_m};$$

avec D_m le m-ème élément de la diagonale de D et g_m la dérivée de $l_m C g l_m'$ par rapport à θ, à la valeur $\hat{\theta}$.

D_m est issu de la décomposition spectrale de $L\hat{C}L' = P'DP$, où P est une matrice orthogonale de vecteurs propres et D est une matrice diagonale de valeurs propres, tous deux de dimensions $q \times q$. l_m est la m-ème ligne de PL.

$C = (X'V^{-1}X)^{-}$, où $()^{-}$ est l'inverse généralisée de la matrice entre parenthèses et C correspond à la matrice de variance/covariance des paramètres fixes et θ est un vecteur des paramètres inconnus des composantes de la variance (partie aléatoire et résiduelle du modèle mixte).

Tests de comparaisons multiples (uniquement pour l'ANOVA sur mesures répétées)

Les modèles mixtes permettent de faire des tests de comparaisons multiples dont le but est de vérifier si les différentes modalités d'un facteur sont significativement différentes ou non.

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.

OK : cliquez sur ce bouton pour lancer les calculs.

Annuler : cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

Aide : cliquez sur ce bouton pour afficher l'aide.

 : cliquez sur ce bouton pour rétablir les options par défaut.

 : cliquez sur ce bouton pour effacer les sélections de données.

 : cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.

Onglet **Général**:

Y / Variables dépendantes :

Quantitatives : sélectionnez la ou les variables réponse que vous souhaitez modéliser. Si plusieurs variables sont sélectionnées, XLSTAT fera les calculs pour chacune des variables indépendamment. Si des en-têtes de colonnes ont été sélectionnés, veuillez vérifier que l'option « Libellés des variables » est activée.

X / Variables explicatives :

Quantitatives : activez cette option si vous voulez inclure une ou plusieurs variables explicatives quantitatives dans le modèle. Les données sélectionnées doivent être de type numérique. Si le libellé des variables a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

Qualitatives : sélectionnez la ou les variables explicatives qualitatives (les facteurs) sur la feuille Excel. Les données sélectionnées peuvent être de tout type, mais les données numériques sont automatiquement considérées comme nominales. Si le libellé des variables a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des variables : activez cette option si la première ligne des données sélectionnées (variables dépendantes et explicatives, libellés des observations, poids) contient un libellé.

Poids des observations : activez cette option si vous voulez pondérer les observations. Si vous n'activez pas cette option, les poids seront tous considérés comme valant 1. Les poids doivent être impérativement supérieurs ou égaux à 0. Un poids de 2 est équivalent à répéter

deux fois la même observation. Si un en-tête de colonne a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

Poids dans la régression : activez cette option si vous voulez pondérer l'équation du modèle. Si vous n'activez pas cette option, les poids seront tous considérés comme valant 1. Les poids doivent être impérativement supérieurs ou égaux à 0. Si un en-tête de colonne a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

Onglet **Options**:

Constante fixée : activez cette option pour fixer la constante du modèle de régression à une valeur que vous devez ensuite saisir (0 par défaut).

Tolérance : activez cette option pour permettre à l'algorithme de ne pas prendre en compte les variables qui seraient soit constantes soit trop corrélées avec d'autres variables déjà utilisées dans le modèle (0.0001 par défaut).

Interactions / Niveau : activez cette option pour inclure des interactions dans le modèle puis entrez le niveau maximum d'interaction (valeur comprise entre 1 et 4).

Intervalle de confiance (%) : entrez l'étendue en pourcentage de l'intervalle de confiance à utiliser pour les différents tests, et pour le calcul des intervalles de confiance autour des paramètres et des prédictions. Valeur par défaut : 95.

Contraintes : des détails sur les différentes options sont disponibles dans la section description.

a1 = 0 : choisissez cette option pour que le paramètre de la première modalité de chaque facteur soit fixé à 0.

an = 0 : choisissez cette option pour que le paramètre de la dernière modalité de chaque facteur soit fixé à 0.

Mesures répétées : activez cette option si vous désirez effectuer une ANOVA avec mesures répétées en utilisant les modèles mixtes.

Structure de covariance : différentes structures sont possibles pour la matrice de covariance des erreurs. On peut choisir parmi : Autoregressive(1), compound symmetry, Toeplitz, Toeplitz(q), Unstructured et Variance Components. Voir la description de la méthode pour le détail.

Facteur aléatoire : (uniquement pour les modèles mixtes): activez cette option si vous désirez utiliser un modèle avec des facteurs aléatoires.

Structure de covariance : différentes structures sont possibles pour la matrice de covariance des coefficients associés aux facteurs aléatoires. On peut choisir parmi : Autoregressive(1), compound symmetry, Unstructured et Variance Components. Voir la description de la méthode pour le détail.

t-tests Satterthwaite : Activez cette option si vous souhaitez calculer les tests t pour les coefficients fixes *Beta* à l'aide de la formule de Satterthwaite pour les degrés de liberté du dénominateur. Les tests F des effets fixes seront également calculés selon cette méthode. Les modèles mixtes avec des jeux de données non équilibrés sont automatiquement calculés à l'aide de l'approximation de Satterthwaite.

Valeurs initiales de Newton-Raphson : Choisissez la méthode pour définir les valeurs initiales de l'algorithme itératif de Newton-Raphson. Utilisez l'option OLS pour spécifier les valeurs initiales des moindres carrés ordinaires ou MIVQUE0 (Estimation de la variance quadratique non biaisée à la variance minimale). Lorsque la structure de covariance est basique, les estimations MIVQUEO sont les estimations REML. Pour une structure de covariance plus complexe, MIVQUE0 pourrait être choisi de manière à être aussi proche que possible des valeurs de la population afin que la routine d'optimisation puisse converger vers des estimations raisonnables.

Méthode d'estimation : deux méthodes sont disponibles et sont toutes deux basées sur le maximum de vraisemblance. La méthode REML (par défaut) et la méthode ML. Voir la description de la méthode pour le détail.

Onglet **Données manquantes** :

Supprimer les observations : activez cette option pour supprimer les observations comportant des données manquantes.

Estimer les données manquantes : activez cette option pour estimer les données manquantes avant le début des calculs.

- **Moyenne ou mode** : activez cette option pour estimer les données manquantes en utilisant la moyenne (variables quantitatives) ou le mode (variables qualitatives) pour les variables correspondantes.
- **Plus proche voisin** : activez cette option pour estimer les données manquantes d'une observation en recherchant le plus proche voisin de l'observation.

Onglet **Sorties**:

Général :

Statistiques descriptives : activez cette option pour afficher les statistiques descriptives pour les variables sélectionnées.

Corrélations : activez cette option pour afficher la matrice de corrélation pour les variables quantitatives (dépendantes et explicatives).

Coefficients d'ajustement : activez cette option pour afficher le tableau des coefficients d'ajustement du modèle.

Paramètres de covariance : activez cette option pour afficher le tableau des estimations des paramètres de covariance.

Test du rapport des vraisemblances : activez cette option pour afficher le tableau du test du rapport de vraisemblances.

Coefficients des effets fixes : activez cette option pour afficher le tableau des coefficients associés aux effets fixes.

Coefficients des effets aléatoires (uniquement pour les modèles mixtes)**:** activez cette option pour afficher le tableau des coefficients associés aux effets aléatoires.

Tests de type III des effets fixes : activez cette option pour afficher les résultats du test de type III sur chaque effet fixe.

Tests de type I des effets fixes : activez cette option pour afficher les résultats du test de type I sur chaque effet fixe.

Tests de type II des effets fixes : activez cette option pour afficher les résultats du test de type II sur chaque effet fixe.

Matrice R : activez cette option pour afficher la matrice de covariance des erreurs R pour le premier sujet.

Matrice G (uniquement pour les modèles mixtes): activez cette option pour afficher la matrice de covariance des effets aléatoires G.

Résidus :

Prédictions et résidus : activez cette option pour afficher les prédictions et les résidus pour l'ensemble des observations.

- **Résidus bruts** : activez cette option pour calculer et afficher les résidus bruts dans le tableau des prédictions et résidus.
- **Résidus studentisés** : activez cette option pour calculer et afficher les résidus studentisés dans le tableau des prédictions et résidus.
- **Résidus de Pearson** : activez cette option pour calculer et afficher les résidus de Pearson dans le tableau des prédictions et résidus.

Comparaisons :

Comparaisons multiples :

Des informations sur les tests de comparaisons multiples sont disponibles dans la section description.

Appliquer à tous les facteurs : activez cette option pour calculer les tests sélectionnés pour tous les facteurs.

Utiliser les moyennes estimées : activez cette option pour calculer les moyennes en utilisant le modèle. Si cette option n'est pas activée, les moyennes sont estimées à partir des données.

Trier en ordre croissant : activez cette option pour trier les modalités comparées en ordre croissant, le critère de tri étant leur moyenne respective. Si cette option n'est pas activée, le tri est décroissant.

Comparaison par paires : activez cette option puis choisissez les méthodes de comparaison.

Boîte de dialogue facteurs et interactions

Une fois la première boîte de dialogue passée, une seconde boîte apparaît. Celle-ci est intitulée facteurs et interactions, son aspect dépendra des options sélectionnées dans la première boîte de dialogue.

Si les mesures répétées ont été sélectionnées ou que l'on effectue une ANOVA avec mesures répétées, il faut sélectionner les facteurs et interactions (effets fixes), un facteur répété et un facteur sujet.

Si on veut prendre en compte un effet aléatoire, alors il faut sélectionner des facteurs fixes et des facteurs aléatoires.

Si on veut traiter des mesures répétées ainsi que des facteurs aléatoires, il faut sélectionner des facteurs fixes, des facteurs aléatoires, un facteur répété et un facteur sujet.

Chaque facteur ne peut être sélectionné que dans une seule colonne. Les facteurs répété et sujet doivent être des variables qualitatives simples.

Résultats

Statistiques descriptives : les tableaux de statistiques descriptives présentent pour toutes les variables sélectionnées des statistiques simples. Pour les variables dépendantes (en bleu) et les variables explicatives quantitatives, sont affichés le nombre d'observations, le nombre de données manquantes, le nombre de données non manquantes, la moyenne, et l'écart-type (non biaisé). Pour les variables explicatives qualitatives sont affichés le nom des différentes modalités ainsi que leur fréquence respective.

Matrice de corrélation : ce tableau est affiché afin de vous permettre d'avoir un aperçu des corrélations entre les différentes variables sélectionnées.

Coefficients d'ajustement : dans ce tableau sont affichées les statistiques relatives à l'ajustement du modèle de régression :

- **Observations** : le nombre d'observations prises en compte dans les calculs. Dans les formules présentées ci-dessous n désigne le nombre d'observations.
- **Somme des poids** : la somme des poids des observations prises en compte dans les calculs. Dans les formules présentées ci-dessous W désigne la somme des poids.
- **AIC** : le critère d'information d'Akaike (Akaike's Information Criterion) est défini par

$$AIC = -2l(\theta) + 2d$$

où l est la fonction de vraisemblance estimée et d est égal au nombre de paramètres estimés. Ce critère proposé par Akaike (1973) dérive de la théorie de l'information, et s'appuie sur la mesure de Kullback et Leibler (1951). C'est un critère de sélection de modèles qui pénalise les modèles pour lesquels l'ajout de nouvelles variables explicatives n'apporte pas suffisamment d'information au modèle. On cherche à minimiser le critère AIC.

- **AICC** : ce critère (Hurvich et Tsai, 1989) issu de l'AIC est défini par :

$$AICC = -2l(\theta) + 2dn/(n - d - 1)$$

- **SBC** : le critère Bayésien de Schwarz (Schwarz's Bayesian Criterion) est défini par

$$SBC = -2l(\theta) + d\ln(n)$$

Ce critère proposé par Schwarz (1978) est proche du critère AIC, et comme ce dernier on cherche à le minimiser. Il est aussi appelé BIC.

- **CAIC** : ce critère (Bodzogan, 1987) est défini par :

$$CAIC = -2l(\theta) + d(\ln(n) + 1)$$

- **Itérations** : nombre d'itérations nécessaires à la convergence de l'algorithme de Newton-Raphson.
- **Paramètres de covariance** : nombre de paramètres à estimer dans la matrice V.
- **Nombre d'effets fixes :** nombre d'effets fixes sélectionnés.
- **Nombre d'effets aléatoires :** nombre d'effets aléatoires sélectionnés.

Paramètres de covariance – Facteur répété : dans ce tableau sont affichés les paramètres associés à la matrice de covariance des erreurs R. Pour chaque paramètre, l'écart type, la valeur de la statistique Z ainsi que la p-valeur associée sont obtenus.

Paramètres de covariance – Facteur aléatoire (uniquement pour les modèles mixtes)**:** dans ce tableau sont affichés les paramètres associés à la matrice de covariance des coefficients des effets aléatoires G. Pour chaque paramètre, l'écart type, la valeur de la statistique Z ainsi que la p-valeur associée sont obtenus.

Test du rapport de vraisemblance : dans ce tableau sont affichés les résultats du test de comparaison des vraisemblances associées, d'une part, au modèle sans variables explicatives et, d'autre part, au modèle sélectionné. Le rapport de vraisemblance, la statistique du chi² ainsi que la p-valeur sont affichés.

Paramètres du modèle : ce tableau rassemble les coefficients associés aux effets fixes du modèle. Pour chaque variable, l'écart-type, la statistique t, la p-valeur, ainsi que l'intervalle de confiance sont calculés.

Coefficients des effets aléatoires (uniquement pour les modèles mixtes)**:** ce tableau rassemble les coefficients associés aux effets aléatoires du modèle. Pour chaque variable, l'écart-type, le nombre de degrés de liberté, la statistique t, la p-valeur, ainsi que l'intervalle de confiance sont calculés.

Si les options tests de type I des effets fixes, et tests de type III des effets fixes sont activées, les tableaux correspondants sont affichés.

Le tableau des **tests de type I des effets fixes** permet de visualiser l'influence de l'ajout progressif des variables explicatives sur l'ajustement du modèle, au sens du F de Fisher, ou de la probabilité associée au F de Fisher. Plus la probabilité est faible, plus la contribution de la variable au modèle est importante, toutes les autres variables étant déjà dans le modèle. Remarque : l'ordre de sélection des variables dans le modèle influe sur les valeurs obtenues.

Le tableau des **tests de type III des effets fixes** permet de visualiser l'influence du retrait d'une variable explicative sur l'ajustement du modèle, toutes les autres variables étant conservées, au sens du F de Fisher, ou de la probabilité associée au F de Fisher. Plus la probabilité est faible, plus la contribution de la variable au modèle est importante, toutes les autres variables étant déjà dans le modèle.

Dans le tableau des **prédictions et résidus** sont donnés pour chaque observation, son poids, la valeur de la variable dépendante, la prédiction du modèle, les intervalles de confiance, ainsi que les résidus. Plusieurs types de résidus sont affichés :

- les résidus bruts :

$$r_i = y_i - x_i' \hat{\beta}$$

- les résidus studentisés :

$$r_i^{stud} = \frac{r_i}{\sqrt{var(r_i)}}$$

- les résidus de Pearson :

$$r_i^{stud} = \frac{r_i}{\sqrt{var(y_i)}}$$

Si au moins un effet aléatoire est sélectionné, on a alors :

- les résidus conditionnels :

$$r_i^{cond} = r_i - z_i' \hat{\gamma}$$

- les résidus conditionnels studentisés :

$$r_i^{cond/stud} = \frac{r_i^{cond}}{\sqrt{var(r_i^{cond})}}$$

- les résidus conditionnels de Pearson :

$$r_i^{cond/pearson} = \frac{r_i}{\sqrt{var(y_i)}}$$

Si des tests de comparaison multiples ont été demandés, les résultats correspondant sont ensuite affichés.

Exemple

Un exemple d'ANOVA avec mesures répétées est disponible sur le Centre d'aide XLSTAT à l'adresse suivante :

http://www.xlstat.com/demo-anorepf.htm

Un exemple de modèle à facteurs aléatoires est disponible sur le Centre d'aide XLSTAT à l'adresse suivante :

http://www.xlstat.com/demo-mixedf.htm

Bibliographie

Akaike H. (1973). Information theory and the extension of the maximum likelihood principle. In: Second International Symposium on Information Theory. (Eds: V.N. Petrov and F. Csaki). Academiai Kiadó, Budapest. 267-281.

Bodzogan, H. (1987). Model selection and Akaike's Information Criterion (AIC)! The General Theory and its Analytical Extensions. *Psychometrika*, **52**, 345-370.

Dempster A.P. (1969). Elements of Continuous Multivariate Analysis. Addison-Wesley, Reading.

Goodnight, J. H. (1979). A Tutorial on the Sweep Operator, *American Statistician*, **33**, 149–158.

Hurvich, C. M. and Tsai, C.-L. (1989). Regression and Time Series Model Selection in Small Samples, *Biometrika*, **76**, 297–307.

Kullback S. and Leibler R. A. (1951). On information and sufficiency. *Annals of Mathematical Statistics*, **22**, 79-86.

Rao, C. R. (1972). Estimation of Variance and Covariance Components in Linear Models, *Journal of the American Statistical Association*, **67**, 112–115.

Sahai H. and Ageel M.I. (2000). The Analysis of Variance. Birkhaüser, Boston.

Schwarz, G. (1978). Estimating the Dimension of a Model, *Annals of Statistics*, **6**, 461–464.

Searle, S. R., Casella, G., and McCulloch, C. E. (1992). Variance Components. John Wiley & Sons, New York.

Wolfinger, R. D. (1993). Covariance Structure Selection in General Mixed Models, *Communications in Statistics, Simulation and Computation*, **22(4)** , 1079–1106.

Wolfinger, R. D., Tobias, R. D., and Sall, J. (1994). Computing Gaussian Likelihoods and Their Derivatives for General Linear Mixed Models, *SIAM Journal on Scientific Computing*, **15(6)**, 1294–1310.

Elizabeth Eskow and Bobby Schnabel (1991). Algorithm 695: software for a new modified Cholesky factorization, *ACM Trans. Math. Softw.*, **17**, 306-312.

Hrong-Tai Fai, Alex & L. Cornelius, Paul. (1996). Approximate F-tests of multiple degree of freedom hypotheses in generalized least squares analyses of unbalanced split-plot experiments, *Journal of Statistical Computation and Simulation*, **54**, 363-378.

MANOVA

Utilisez ce module pour faire de la MANOVA (Analyse de variance multivariée) à deux ou plus de facteurs, équilibrée ou déséquilibrée. Des options vous permettent de choisir le niveau de confiance et de tenir compte des interactions entre les facteurs. Des tests multivariés peuvent être calculés.

Dans cette section :

Description

L'analyse de variance multivariée utilise le même cadre conceptuel que l'ANOVA. La différence principale vient de la nature des variables dépendantes : on peut en considérer plusieurs en même temps. Dans le cadre de la MANOVA, les variables explicatives sont souvent appelées facteurs.

L'analyse de la variance multivariée est donc une extension de l'ANOVA dans laquelle les effets des facteurs sont évalués sur une combinaison de plusieurs variables réponses.

L'avantage de l'utilisation d'une MANOVA au lieu de plusieurs ANOVA simultanée réside dans le fait qu'elle prend en compte la corrélation entre les variables réponses et permet ainsi une meilleure utilisation des informations provenant des données.

Ainsi, la MANOVA teste la présence de différences significatives parmi les combinaisons de niveaux de facteurs sur plusieurs variables réponses. Avec une MANOVA, on est donc capable de tester conjointement toutes les hypothèses que testent une ANOVA et on a plus de chances d'observer les différences entre les niveaux de facteurs. De plus, faire plusieurs ANOVA au lieu d'une MANOVA augmente l'erreur de type I c'est à dire la probabilité de rejeter à tort l'hypothèse H0.

Enfin, plusieurs ANOVA séparées ne prennent pas en compte la covariation entre variables réponses tandis que la MANOVA n'est pas seulement sensible aux différences de moyenne entre niveaux de facteurs mais également à la covariance entre variables explicatives. Quand

ces variables sont toutes étudiées ensemble, il y a plus de chances de détecter une possible corrélation entre certaines variables. Ce n'est pas le cas avec une ANOVA qui ne prend en compte qu'une seule variable réponse.

Prenons comme exemple illustratif une MANOVA à deux facteurs A et B, le modèle de la MANOVA s'écrit de la manière suivante :

$$y_{ijk} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + \epsilon_{ijk} \quad (1)$$

où y_{ijk} est la kème observation du ième niveau du facteur A et du jème niveau du facteur B, et ϵ est l'erreur du modèle.

Les hypothèses utilisées en MANOVA sont identiques à celles de la régression linéaire : les erreurs ϵ suivent une même loi normale $N(0, s)$ et sont indépendantes.

L'écriture du modèle complétée par cette hypothèse a pour conséquence que, dans le cadre du modèle de régression linéaire, les y_{ijk} sont des réalisations de variables aléatoires de moyenne μ et de variance σ^2.

Si l'on souhaite utiliser les différents tests proposés dans les résultats de la régression linéaire il est recommandé de vérifier *a posteriori* que les hypothèses sous-jacentes sont bien vérifiées. La normalité des résidus peut être vérifiée en analysant certains graphiques ou en utilisant un test de normalité. L'indépendance des résidus peut être vérifiée en analysant certains graphiques ou en utilisant le test de Durbin-Watson.

Interactions

On désigne par interaction un facteur artificiel (non mesuré) reflétant l'interaction entre au moins deux facteurs mesurés. Par exemple, si on applique un traitement à une plante, et que les essais sont réalisés sous deux intensités lumineuses différentes, on pourra inclure dans le modèle un facteur d'interaction traitement*lumière qui permettra d'identifier une éventuelle interaction entre les deux facteurs. S'il y a une interaction entre les deux facteurs, on observera sur les plantes un effet significativement plus important lorsque la lumière est forte et que le traitement est de type 2, alors que l'effet est moyen pour les couples (lumière faible, traitement 2) et (lumière forte, traitement 1).

Pour faire un parallèle avec la régression linéaire, les interactions sont équivalentes à des produits entre les valeurs explicatives continues, bien qu'ici l'obtention des interactions nécessite plus qu'une simple multiplication entre deux variables. Néanmoins la notation utilisée pour représenter l'interaction entre le facteur A et le facteur B est A*B.

XLSTAT permet de facilement définir les niveaux d'interactions à prendre en compte dans le modèle.

MANOVA équilibrée et déséquilibrée

On parle de MANOVA équilibrée lorsque les effectifs des modalités sont égaux pour l'ensemble des combinaisons de facteurs. Lorsque les effectifs de toutes les modalités de l'une des combinaisons de facteurs ne sont pas égaux, alors la MANOVA est dite déséquilibrée.

XLSTAT permet de traiter les deux cas.

Contraintes

Au cours des calculs, chaque facteur est décomposé en une sous-matrice contenant autant de colonnes qu'il y a de modalités dans le facteur. Typiquement, il s'agit d'un tableau disjonctif complet. Cette décomposition pose néanmoins un problème : s'il y a g modalités, le rang de cette sous- matrice n'est pas g mais $g-1$. Cela entraîne la nécessité de supprimer l'une des colonnes de la sous-matrice, et éventuellement de transformer les autres colonnes. La stratégie adoptée dans XLSTAT est la suivante :

a1=0 : le paramètre correspondant à la première modalité est nul. Ce choix permet d'imposer que l'effet de la première modalité correspond à un standard. Dans ce cas, la constante du modèle est égale à la moyenne de la variable dépendante pour chaque premier niveau des différents facteurs.

De plus, le nombre d'observations doit être au moins égal à la somme du nombre de variables dépendantes et du nombre de facteurs et d'interactions dans le modèle (+1).

Tests multivariés

L'une des applications principales de la MANOVA sont les tests multivariés dont le but est de vérifier si les paramètres correspondant aux différentes modalités d'un facteur sont significativement différents ou non. Par exemple, dans le cas où quatre traitements sont appliqués à des plantes, on veut savoir non seulement si les traitements ont un effet significatif, mais aussi si les traitements ont un effet différent.

De nombreux tests ont été proposés pour comparer les moyennes des modalités. La majorité de ces tests s'appuie sur la relation entre la matrice d'erreur E et la matrice qui symbolise l'hypothèse testée dans le modèle H, c'est-à-dire les valeurs propres de la matrice $E^{-1}H$. XLSTAT propose les principaux tests parmi lesquels :

Test de Wilks Lambda : le test de rapport de vraisemblance plus connu sous le nom de test de Wilks Lambda (Wilks-1932) est égal à :

$$Lambda = \prod_{i=1}^{m} \frac{1}{1+\lambda_i}$$

Cette valeur peut être interprétée comme le pourcentage de variabilité non expliquée par l'effet étudié. Ce test est le plus utilisé, il reste interprétable pour des données équilibrées et des données déséquilibrées.

Si E est petit par rapport à H, Lambda sera proche de 0 sinon Lambda sera proche de 1.

L'hypothèse nulle est rejetée quand Lambda est petit.

Test de la trace de Hotelling-Lawley :

$$T_{HL} = \sum_{i=1}^{m} \lambda_i = Trace(E^{-1}H)$$

Plus H est grand comparé à E, plus la trace est grande. Dans ce cas, on rejette l'hypothèse nulle pour des grandes valeurs de T_{HL}. Ce test est utile dans le cas où toutes les variables qualitatives ont toutes exactement deux niveaux. Dans ce cas, il est robuste.

Test de la trace de Pillai :

$$T_P = \sum_{i=1}^{m} \frac{\lambda_i}{1+\lambda_i} = \mathrm{Trace}((E+H)^{-1}H)$$

Comme le test de la trace de Hotelling-Lawley, l'hypothèse nulle est rejetée pour de grandes valeurs de T_P. Ce test est efficace dans le cas où on a égalité des échantillons.

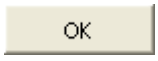
Test de la plus grande racine de Roy :

$$\lambda_{max} = \max_{1 \leq i \leq m} \lambda_i$$

La p-valeur calculée pour ce test est toujours plus petite que celle des autres. Le test de Roy est un test puissant mais pas robuste. Pour cette raison, il est déconseillé.

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les

variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.

Onglet **Général**:

Y / Variables dépendantes :

Sélectionnez la ou les variables réponse que vous souhaitez modéliser. Si plusieurs variables sont sélectionnées, XLSTAT fera les calculs pour chacune des variables indépendamment. Si des en-têtes de colonnes ont été sélectionnés, veuillez vérifier que l'option « Libellés des variables » est activée.

X / Variables explicatives :

Sélectionnez la ou les variables explicatives qualitatives (les facteurs) sur la feuille Excel. Les données sélectionnées peuvent être de tout type, mais les données numériques sont automatiquement considérées comme nominales. Si le libellé des variables a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des colonnes : activez cette option si la première ligne des données sélectionnées (variables dépendantes et explicatives, libellés des observations, poids) contient un libellé.

Poids des observations : activez cette option si vous voulez pondérer les observations. Si vous n'activez pas cette option, les poids seront tous considérés comme valant 1. Les poids doivent être impérativement supérieurs ou égaux à 0. Un poids de 2 est équivalent à répéter deux fois la même observation. Si un en-tête de colonne a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

Onglet **Options**:

Intervalle de confiance (%) : entrez le niveau de signification pour les différents tests calculés.

Interactions / Niveau : activez cette option pour inclure des interactions dans le modèle puis entrez le niveau maximum d'interaction (valeur comprise entre 1 et 4).

Onglet **Données manquantes** :

Supprimer les observations : activez cette option pour supprimer les observations comportant des données manquantes.

Ne pas autoriser les données manquantes : activez cette option ne pas accepter que son jeu de données contienne des données manquantes.

Onglet **Sorties**:

Statistiques descriptives : activez cette option pour afficher les statistiques descriptives pour les variables sélectionnées.

Moyennes par niveau de facteur : activez cette option pour afficher les moyennes de chaque niveau de facteurs.

Matrices SSCP : activez cette option pour afficher les matrices des sommes de carrés et produits croisés pour chaque variable explicative et, éventuellement, chaque interaction.

Valeurs propres : activez cette option pour afficher les valeurs propres calculées sur les matrices SSCP pour chaque facteur et, éventuellement, chaque interaction.

Résultats de tests : activez cette option pour afficher les résultats des tests statistiques.

Wilks : activez cette option si vous voulez les résultats du test de Wilks Lambda.

Hotelling-Lawley : activez cette option si vous voulez les résultats du test de la trace de Hotelling-Lawley.

Pillai : activez cette option si vous voulez les résultats du test de la trace de Pillai.

Roy : activez cette option si vous voulez les résultats du test de la plus grande racine de Roy.

Onglet **Graphiques**:

Graphique des moyennes : activez cette option pour afficher les moyennes de chaque niveau de facteurs sous forme d'histogramme.

Résultats

Statistiques descriptives : les tableaux de statistiques descriptives présentent pour toutes les variables sélectionnées des statistiques simples. Pour les variables dépendantes et les variables explicatives quantitatives, sont affichés le nombre d'observations, le nombre de données manquantes, le nombre de données non manquantes, la moyenne, et l'écart-type (non biaisé). Pour les variables explicatives qualitatives sont affichés le nom des différentes modalités ainsi que leur fréquence respective.

Moyennes par niveau de facteur : ce tableau présente pour chaque niveau des facteurs la moyenne de leur valeur pour chaque variable quantitative.

Matrices SSCP : ces tableaux sont affichés afin de vous permettre d'avoir un aperçu de l'effet des facteurs et interactions du modèle.

Test de Wilks (approximation de Rao) : ce tableau fournit les résultats du test du Lambda de Wilks qui teste l'hypothèse d'égalité des vecteurs moyens des différents niveaux. Lorsqu'il y a deux niveaux le test est équivalent au test de Fisher. Si le nombre de niveaux est inférieur ou égal à trois, le test est exact. L'approximation de Rao est nécessaire à partir de quatre niveaux pour obtenir une statistique approximativement distribuée suivant une loi de Fisher.

Test de Hotelling-Lawley : ce tableau fournit les résultats du test de la trace de Hotelling-Lawley qui teste l'hypothèse d'égalité des vecteurs moyens des différents niveaux. Il est moins utilisé que le test du Lambda de Wilks et utilise aussi la loi de distribution de Fisher pour le calcul des p-values.

Test de Pillai : ce tableau fournit les résultats du test de Pillai qui teste l'hypothèse d'égalité des vecteurs moyens des différents niveaux. Il est moins utilisé que le test du Lambda de Wilks et utilise aussi la loi de distribution de Fisher pour le calcul des p-values.

Test de Roy : ce tableau fournit les résultats du test de la plus grande racine de Roy qui teste l'hypothèse d'égalité des vecteurs moyens des différents niveaux. Il est moins utilisé que le test du Lambda de Wilks et utilise aussi la loi de distribution de Fisher pour le calcul des p-values.

Exemple

Un exemple de MANOVA à un facteur est disponible sur le Centre d'aide XLSTAT à l'adresse suivante :

http://www.xlstat.com/demo-manof.htm

Bibliographie

Barker H. R. & Barker B. M. (1984). *Multivariate analysis of variance (MANOVA): a practical guide to its use in scientific decision-making*., University of Alabama Press.

Gentle, J. E., Härdle W. K. & Mori Y. (2012). *Handbook of computational statistics: concepts and methods*., Springer Science & Business Media.

Hand D ;J. & Taylor C.C. (1987). *Multivariate analysis of variance and repeated measures: a practical approach for behavioural scientists.*, Chapman & Hall.

Taylor, A. (2011). Multivariate Analyses of variance with manova and GLM. psy.mq.edu.au/psystat/documents/Multivariate.pdf

Zetterberg, P. (2013). Effects of unbalancedness and heteroscedasticity on two way MANOVA., Department of statistics, Stockholm University.

- chm-mapid: 130

Régression logistique

Utilisez la régression logistique pour modéliser une variable qualitative binaire (2 modalités), ordinale (plus de deux modalités ordonnées) ou polytomique (plus de deux modalités) en fonction de variables explicatives quantitatives ou qualitatives.

Dans cette section :

Description

La régression logistique est une méthode très utilisée car elle permet de modéliser des variables binomiales (typiquement binaires), multinomiales (variables qualitatives à plus de deux modalités) ou ordinales (variables qualitatives dont les modalités peuvent être ordonnées). Elle est très utilisée dans le domaine médical (guérison ou non d'un patient), en sociologie, en épidémiologie, en marketing quantitatif (achat ou non de produits ou services suite à une action) et en finance pour la modélisation de risques (scoring).

Le principe du modèle de la régression logistique est d'expliquer la survenance ou non d'un événement (la variable dépendante notée Y) par le niveau de variables explicatives (notées X). Par exemple, dans le domaine médical, on cherche à évaluer à partir de quelle dose d'un médicament, un patient sera guéri.

Cas de la régression logistique pour des variables réponse binomiales

La régression logistique et la régression linéaire appartiennent à la même famille des modèles GLM (*Generalized Linear Models*) : dans les deux cas, on relie un événement à une combinaison linéaire de variables explicatives.

Dans le cas de la régression linéaire ordinaire, la variable dépendante Y suit une loi normale $N(\mu, \sigma)$ où μ est une fonction linéaire des variables explicatives. Pour la régression logistique binomiale, la variable dépendante, aussi appelée variable réponse, suit une loi de Bernoulli de paramètre p (p étant la probabilité pour que l'événement se produise), lorsque l'expérience est répétée une fois, ou une loi Binomiale(n, p) si l'expérience est répétée n fois (par exemple la même dose est essayée sur n insectes). Dans le cas de la régression logistique, le paramètre de probabilité p est une fonction d'une combinaison linéaire des variables explicatives X.

XLSTAT nomme "binaire" le cas où la variable réponse peut prendre 2 valeurs (correspondant à un tirage de Bernoulli), et "somme de binaires" le cas où la variable réponse est le comptage du

nombre de fois où l'événement d'intérêt s'est produit.

Les fonctions les plus couramment utilisées pour relier la probabilité p aux variables explicatives sont la fonction logistique (on parle alors de modèle Logit) et la fonction de répartition de la loi normale standard (on parle alors de modèle Probit). Ces deux fonctions sont parfaitement symétriques et sigmoïdes. XLSTAT propose deux autres fonctions : la fonction Log-log complémentaire qui n'est plus symétrique car concentrée sur l'asymptote supérieure, et la fonction de Gompertz qui est au contraire plus concentrée sur l'axe des abscisses.

L'expression analytique des modèles est donnée ci-dessous :

1. Logit : $p = \frac { \operatorname{exp} (\beta X) } { 1+ \operatorname{exp} (\beta X) } = \frac { 1 } { 1+ \operatorname{exp} (-\beta X) } $
2. Probit : $p = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\beta X} \exp[\frac{-x^2}{2}] dx$
3. Log-log complémentaire : $p = 1 - \exp[-\exp(\beta X)]$
4. Gompertz : $p = exp[-exp(-\beta X)]$

où βX représente la combinaison linéaire des q variables explicatives (constante comprise).

La connaissance de la loi de distribution de l'événement étudié, permet d'écrire la vraisemblance de l'échantillon. Pour estimer les paramètres β du modèle (les coefficients de la fonction linéaire), on cherche à maximiser la fonction de vraisemblance. Contrairement ce qui est le cas pour la régression linéaire, une solution analytique exacte n'existe pas. Il est donc nécessaire d'utiliser un algorithme itératif. XLSTAT utilise un algorithme de Newton-Raphson. L'utilisateur peut modifier s'il le souhaite le nombre maximum d'itérations, seuil de convergence (la précision à atteindre pour la fonction de vraisemblance) ou le temps maximum à allouer à la recherche du maximum.

Dans le cas de la régression logistique binomiale, lorsque l'on connaît la probabilité p associée à une modalité (typiquement 1), on en déduit la probabilité $1 - p$ d'appartenance à l'autre modalité (typiquement 0). La modalité de référence est par défaut la première modalité dans l'ordre alphabétique, soit 0 si on est dans un cas (0/1). Le modèle est calculé pour la modalité qui n'est pas celle de référence, mais il est très simple d'en déduire la probabilité pour la modalité de référence.

Dans la plupart des logiciels, le calcul des intervalles de confiance sur les paramètres est comme pour la régression linéaire basé une hypothèse de normalité des paramètres. XLSTAT propose aussi la méthode alternative LR (*likelihood ratio*) introduite par Venzon et Moolgavkar (1988). Cette méthode est plus fiable car elle ne nécessite pas de supposer la normalité des paramètres ; elle peut néanmoins ralentir les calculs car elle est itérative.

Cas de la régression logistique multinomiale

Le principe de la régression logistique multinomiale est d'expliquer ou de prédire une variable pouvant prendre J valeurs alternatives (les J modalités de la variable), en fonction de variables explicatives. Le cas binomial vu précédemment en est donc un cas particulier.

Dans le cadre du modèle multinomial, une modalité de référence doit être sélectionnée. Dans l'interface de XLSTAT elle est appelée « modalité témoin ». Idéalement, on choisira ce qui correspond à la situation "de base" ou "classique" ou "normale". Les coefficients estimés seront interprétés en fonction de cette modalité de référence. Pour la simplicité de l'écriture, les équations ci-dessous sont écrites en considérant que la première modalité comme modalité de référence.

Le modèle proposé par XLSTAT pour relier la probabilité de survenance d'un événement aux variables explicatives est le modèle logit qui est l'un des quatre modèles proposés pour le cas binomial. L'expression analytique du modèle pour les modalités 2 à J est donnée ci-dessous :

$$\log(\frac{p(Y=j|x_i)}{p(Y=1|x_i)}) = \alpha_j + \beta_j X_i \ , \quad i = 2...j$$

$$p(Y=j|X_i) = \frac{exp(\alpha_j + \beta_j X_i))}{1 + \sum_{k=2}^{J} \exp(\alpha_k + \beta_k X_i)} \ , \quad i = 2...j$$

Pour la modalité 1, on a :

$$p(Y=1|x_i) = \frac{1}{1 + \sum_{k-2}^{J} \exp(\alpha_k + \beta_k X_i)}$$

On peut ainsi obtenir la log-vraisemblance de l'échantillon :

$$l(\alpha, \beta) = \sum_{i=1}^{n} \sum_{j=1}^{J} y_{ij} \log(p(Y=j|x_i))$$

Pour estimer les paramètres α et β du modèle (les coefficients de la fonction linéaire), on cherche à maximiser la fonction de vraisemblance. Contrairement à la régression linéaire, une solution analytique exacte n'existe pas. Il est donc nécessaire d'utiliser un algorithme itératif. XLSTAT utilise un algorithme de Newton-Raphson.

Cas de la régression logistique ordinale

Le principe de la régression logistique ordinale est d'expliquer ou de prédire une variable pouvant prendre J valeurs alternatives ordonnées (seul l'ordre importe, pas les écarts), en fonction de variables explicatives. La régression logistique binomiale est un cas particulier de la régression logistique ordinale, correspondant au cas où $J = 2$.

XLSTAT permet d'utiliser deux modèles alternatifs pour calculer les probabilités d'affectation aux modalités à partir des variables explicatives : le modèle logit et le modèle probit.

Pour le cas du modèle logit, on a :

$$\log(\frac{p(Y \leq j|x_i)}{p(y > j|x_i)}) = \alpha_j + \beta X_i \ , \quad i = 1...J-1$$

On voit donc qu'il existe une constante par modalité de Y mais à la différence du modèle multinomial, on a un unique jeu de coefficients β pour l'ensemble des modalités de Y. La modalité de référence est toujours la plus faible. La probabilité de choisir la modalité j ou une modalité plus petite que celle-ci est donnée par :

$$p(Y \leq j|x_i) = \frac{\exp(\alpha_j + \beta X_i)}{1 + \exp(\alpha_j + \beta X_i)}, \quad i = 1...J-1$$

Pour $j = J$, cette probabilité vaut 1.

La probabilité d'obtenir la modalité j est :

$$p(Y = j|x_i) = p(Y \leq j|x_i) - p(Y \leq j-1|x_i)$$

On peut ainsi obtenir la log-vraisemblance de l'échantillon :

$$l(\alpha, \beta) = \sum_{i=1}^{n} \sum_{j=1}^{J} y_{ij} \log(p(Y \leq j|x_i) - p(Y \leq j-1|x_i))$$

Pour estimer les q paramètres β et les $J-1$ paramètres α_j du modèle (les coefficients de la fonction linéaire), on cherche à maximiser la fonction de vraisemblance. Contrairement à la régression linéaire, une solution analytique exacte n'existe pas. Il est donc nécessaire d'utiliser un algorithme itératif. XLSTAT utilise un algorithme de Newton-Raphson.

Prédictions et analyse de l'influence des observations

Quelque soit le type de variable réponse (binaire, somme(binaires), multinomiale ou ordinale), XLSTAT fourni pour chaque observation la probabilité d'affectation à chacune des modalités possibles.

Dans le cas binomial (binaire ou somme(binaires)), pour un point de séparation (*cutoff*) C donné, typiquement 0.5, si la probabilité pour l'observation i est inférieure à cette valeur seuil, l'observation est considérée comme étant affectée à la classe 0, sinon, elle est affectée à la classe 1. Si un point de séparation différent de 0.5 a été choisi, la probabilité de la modalité qui n'est pas celle de référence (dans le cas binaire, est comparée à la valeur seuil et si elle supérieure, alors on prédit la catégorie 1.

La notion de valeur seuil ne s'applique pas dans le cas ordinal ou multinomial et les observations sont toujours affectées à la modalité pour laquelle la probabilité est maximale.

Si l'option **analyse de significativé** a été activée, XLSTAT propose une analyse de la significativé du choix de la modalité d'affectation. En effet, pour le décideur, il est important de savoir dans quelle mesure le choix d'affectation est entâché d'incertitude. Cette pratique est malheureusement trop peu courante et c'est pour l'encourager et la faciliter que XLSTAT affiche ce résultat.

Dans le cas binomial, cette analyse se déduit aumatiquement des intervalles de confiance des probabilités. Si l'intervalle de confiance autour d'une probabilité n'inclut pas le point de séparation, alors le risque d'erreur est limité à 5% (le pourcentage dépend du choix fait pour la

taille des intervalles de confiance). Pour les cas multinomial et ordinal, les calculs sont plus compliqués et XLSTAT est le seul à les réaliser. Les comparaisons sont réalisées deux à deux.

L'analyse de significativité est affichée sur deux colonnes. La première indique si, dans le cas où la modalité prédite n'est pas celle des données, si ce changement est significatif ou non. La deuxième colonne indique elle, quelque soit la modalité retenue, si la probabilité pour cette modalité est supérieure ou non à celles des autres modalités.

Enfin, dans le cas de variables dépendantes binomiales, si l'option **diagnostics d'influence** a été sélectionnée, XLSTAT affiche le tableau donnant différentes statistiques, notamment recommandées par Pregibon (1981).

Notons $\pi_{i,j}$ la probabilité calculée par le modèle que l'on observe pour l'observation i la $j^{ème}$ modalité.

Pour la régression sur variables binomiales, les indices calculés sont : * Résidu : dans le cas "binaire", le résidu e_i vaut $1-\pi_{i,j}$ où j est la modalité observée. Plus le résidu est proche de 0, meilleure est la prédiction. Plus il est proche de 1, moins la prédiction est bonne. Pour le cas somme(binaires), le résidu correspond à la différence entre le nombre de cas observés et prédits. * Résidu du modèle : ce résidu est donné par : $\tilde{e_i} = e_i / \left[\pi_{i,j}(1-\pi_{i,j})\right]$ * Résidu std. (résidu standardisé) : $z_i = e_i / \sqrt{\pi_{i,j}(1-\pi_{i,j})}$. Ce résidu est aussi désigné comme résidu de Pearson. La somme des carrés de ces résidus donne la statistique d'ajustement du χ^2. * Déviance : la déviance d_i permet de mesurer si une observation influe ou non sur le modèle. Plus la valeur est éloignée de 0, plus l'influence est importante. Le signe indique si la valeur observée est inférieure ou non à la valeur prédite. La somme des déviances au carré donne $-2LL$ (LL désigne la log-vraisemblance). * Leverage : le *leverage* h_i (effet de levier en anglais) permet de repérer des observations pour lesquelles les valeurs observées ne sont pas comme attendues par le modèle et elles sont donc atypiques par rapport à une majorité d'autres observations. * Résidu studentisé : ils sont donnés par $\tilde{d_i} = d_i / \sqrt{1-h_i}$ et combinent les notions de déviance et leverage. Plus ils sont élevés, plus l'observation est suspecte au regard du modèle. * Distance de Cook : la distance de Cook est donnée par $c_i = z_i^2 h_i / (1-h_i)$. Plus elle est élevée, plus l'observation mérite d'être étudiée de près, en ce sens qu'elle est atypique et n'a probablement pas été enregistrée dans les mêmes conditions expérimentales que les autres. * DFBeta : cet indice est calculé pour chaque variable explicative. Il permet de mesurer l'impact de chaque observation sur les coefficients de chacune des varaibles explicatives. Le signe des DFBeta indiquent dans quel sens a lieu l'influence, et plus la valeur est élevée, plus l'observation influe sur la valeur du coefficient.

Observations bien et mal classées, indice GCI et courbe ROC

XLSTAT donne la possibilité d'afficher le tableau de classification (aussi appelé matrice de confusion) qui permet de calculer un pourcentage d'observations bien classées.

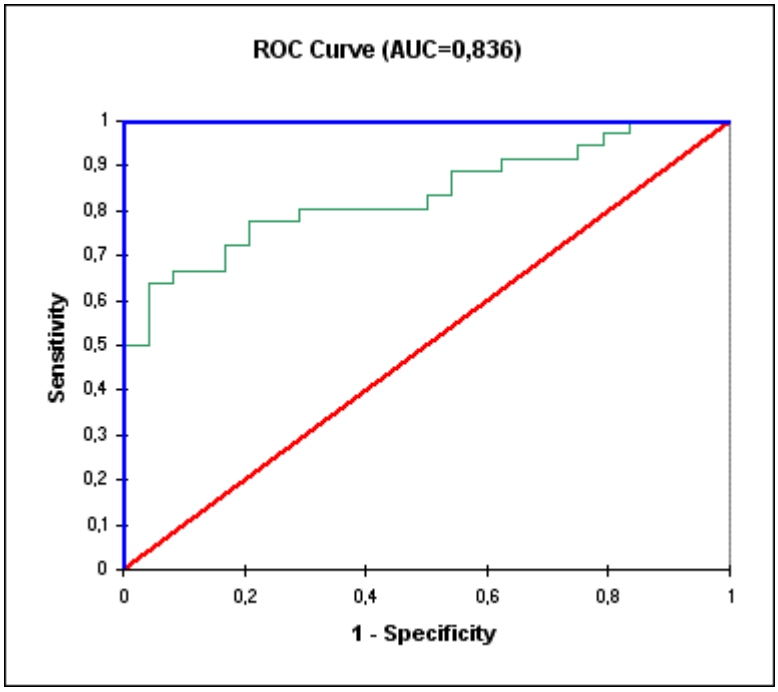
Dans le cas où l'option **analyse de significativé** a été activée, une matrice de confusion avec une colonne supplémentaire "incertain" est ajoutée, afin de compter les observations pour lesquelles la modalité prédite est incertaine (la probabilité la plus élevée n'est pas significativement différente de toutes les autres).

Un tableau synthétise les quatre indices calculés. Soient W la somme des poids des N observations de l'échantillon d'apprentissage, BC le nombre d'individus bien classés, MC le

nombre d'individus mal classés et IC le nombre d'individus incertains. Le tableau présente : * % correct = BC/W : pourcentage d'observations bien classées (vrais positifs) * % incertain IC/W : pourcentage d'observations donc le classement est incertain * % incorrect MC/W : pourcentage d'observations mal classées (faux positifs et faux négatifs cumulés) * $GCI = (BC - MC + IC/2)/W$: le *Goodness of Classification Index* (indice de bonne classification) est un indice dérivé des valeurs précédentes, mis au point par Addinsoft afin d'évaluer simplement et réalistiquement la qualité prédictive d'un modèle de classification. Il est exprimé en %.

Cas particulier des variables binaires : on désigne par sensibilité (*sensivity*) la proportion d'événements positifs bien classés (vrais positifs). La spécificité (*specificity*) correspond à la proportion d'événements négatifs bien classés (vrais négatifs). Si l'on fait varier la probabilité seuil à partir de laquelle on considère qu'un événement doit être considéré comme positif, la sensibilité et la spécificité varient. La courbe des points (1-spécificité, sensibilité) est la courbe ROC. La courbe ROC (*Receiver Operating Characteristics*) permet de visualiser la performance d'un modèle, et de la comparer à celle d'autres modèles. Les termes utilisés viennent de la théorie de détection du signal.

Exemple de courbe ROC : Considérons une variable dépendante binaire indiquant si un client a répondu favorablement à un mailing (oui/non). Sur la figure ci-dessous, la courbe bleue correspond à un cas idéal où les n% de personnes ayant répondu favorablement correspondent aux n% de probabilités les plus élevées. La courbe verte correspond aux résultats d'un modèle bien discriminant. La courbe rouge (première bissectrice) correspond à ce que l'on obtiendrait avec un modèle aléatoire de Bernoulli avec une probabilité de réponse égale à celle observée sur l'échantillon étudié. Un modèle proche de la courbe rouge est donc inefficace puisqu'il n'est pas meilleur qu'un simple tirage au hasard. Un modèle en dessous de cette courbe serait catastrophique car il ferait moins bien que le hasard.



L'aire sous la courbe (ou *Area Under the Curve – AUC*) est un indice synthétique calculé pour les courbes ROC. L'AUC correspond à la probabilité pour qu'un événement positif ait une probabilité donnée par le modèle plus élevée qu'un événement négatif. Pour un modèle idéal, on a AUC=1, pour un modèle aléatoire, on a AUC=0.5. On considère habituellement que le

modèle est bon dès lors que la valeur de l'AUC est supérieure à 0.7. Un modèle bien discriminant doit avoir une AUC entre 0.87 et 0.9. Un modèle ayant une AUC supérieure à 0.9 est excellent.

Problème de séparation

Dans l'exemple ci-dessous, la variable $Traitement$ permet de parfaitement distinguer les cas positifs des cas négatifs.

	Traitement 1	Traitement 2
Réponse +	121	0
Réponse -	0	85

Dans de tels cas, il existe une indétermination sur un ou plusieurs paramètres dont la variance est d'autant plus grande que le seuil de convergence est faible, ce qui empêche de fournir un intervalle de confiance autour du paramètre. Afin de résoudre ce problème et d'obtenir une solution stable, Firth (1993) a proposé d'utiliser une fonction de vraisemblance pénalisée (*penalized likelihood*). XLSTAT propose cette solution en option en s'appuyant sur les résultats fournis par Heinze (2002). Si l'écart type de l'un des paramètres est très élevé par rapport à l'estimation du paramètre, il est conseillé de recommencer les calculs en activant l'option « Firth ».

Depuis la version 2021.3 XLSTAT propose également la pénalisation L2 qui permet également de résoudre ce problème.

Contraintes pour les variables qualitatives

Au cours des calculs, chaque facteur (variable qualitative) est décomposé en une sous-matrice contenant autant de colonnes qu'il y a de modalités dans le facteur. Il s'agit d'un tableau disjonctif complet. Cette décomposition pose néanmoins un problème : s'il y a g modalités, le rang de cette sous- matrice n'est pas g mais $g-1$. Cela entraîne la nécessité de supprimer l'une des colonnes de la sous-matrice, et éventuellement de transformer les autres colonnes. Plusieurs stratégies sont possibles en fonction de l'interprétation que l'on veut ensuite faire :

1) **a1=0** : le paramètre correspondant à la première modalité est nul. Ce choix permet d'imposer que l'effet de la première modalité correspond à un standard. Dans ce cas, la constante du modèle est égale à la moyenne de la variable dépendante pour le groupe 1.

2) **an=0** : le paramètre correspondant à la dernière modalité est nul. Ce choix permet d'imposer que l'effet de la dernière modalité correspond à un standard. Dans ce cas, la constante du modèle est égale à la moyenne de la variable dépendante pour le groupe g.

3) **Somme(ai)=0** : la somme des paramètres est nulle. Ce choix permet d'imposer que la constante du modèle est égale à la moyenne de la variable dépendante lorsque l'ANOVA est équilibrée.

1) **Somme(ni.ai)=0** : la somme pondérée des paramètres est nulle. n_i est le nombre d'observations de la catégorie a_i.

Test de Hosmer-Lemeshow

Le test de Hosmer-Lemeshow est un test permettant de juger la qualité d'ajustement d'un modèle logistique binaire. Une statistique qui suit une distribution du khi² est utilisée.

Le calcul de cette statistique se fait en plusieurs étapes :

- L'échantillon est ordonné de manière décroissante en fonction des probabilités calculées à partir du modèle.
- L'échantillon est découpé en k parties de taille égale.
- La statistique de Hosmer-Lemeshow est calculée grâce à la formule suivante :

$$S_{HL} = \sum_{i=1}^{k} \frac{O(i) - n_i P(i)}{n_i P(i)(1 - P(i))}$$

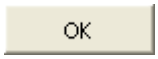
avec n_i taille du groupe i, $O(i)$ nombre de fois où $Y = 1$ dans le groupe i et $P(i)$ moyenne des probabilités obtenues à partir du modèle pour le groupe i.

Cette statistique suit une loi du χ^2 à $k - 2$ degrés de liberté. XLSTAT utilise comme valeur de $k = 10$.

Lorsque cette statistique est grande et que la p-value est petite, la qualité d'ajustement du modèle est mauvaise.

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.

Onglet **Général**:

Variable(s) réponse : sélectionnez la ou les variables réponse que vous souhaitez modéliser. Si plusieurs variables sont sélectionnées, XLSTAT fera les calculs pour chacune des variables indépendamment. Si des en-têtes de colonnes ont été sélectionnés, veuillez vérifier que l'option « Libellés des variables » est activée.

Type de réponse : choisissez le type de variable réponse que vous avez sélectionné :

- **Variable binaire** : si vous sélectionnez cette option, vous devez sélectionner une variable contenant exactement deux valeurs distinctes. Si la variable est constituée de 0 et de 1, XLSTAT fera en sorte que les probabilités élevées du modèle correspondent à la catégorie 1, et que les probabilités faibles correspondent à la catégorie 0. Si la variable comprend deux autres valeurs (par exemple Oui / Non), à la première catégorie rencontrée correspondront les faibles probabilités et à la seconde les probabilités élevées.

- **Somme(binaires)** : si votre variable réponse correspond à une somme de variables binaires, elle doit être de type numérique et contenir le nombre d'événements positifs (événement 1) parmi tous ceux observés. La variable correspondant au nombre total d'événements observés pour cette observation (événements 1 et 0 combinés) doit alors être sélectionnée dans le champ « poids des observations ». Ce cas correspond par exemple à une expérience où l'on administre une dose D d'un médicament (D est la variable explicative) à 50 patients (50 est la valeur du poids des observations), et où l'on observe que 40 sont guéris sous l'effet de la dose (40 correspond à la valeur de la variable réponse).

- **Multinomiale** : si votre variable comporte plus de deux modalités, un modèle logit multinomial est alors estimé. Un nouveau champ appelé « modalité témoin » apparaît, celui-ci permet de choisir la modalité utilisée comme référence.

- **Ordinale** : si votre variable comporte des modalités dont l'ordre a une signification, un modèle logit ordinal est alors estimé. La modalité de référence utilisée est la plus faible. Les données utilisées doivent être numériques avec un nombre de modalités limité.

Variables explicatives :

Quantitatives : activez cette option si vous voulez inclure une ou plusieurs variables explicatives quantitatives dans le modèle. Sélectionnez alors la ou les variables correspondantes sur la feuille Excel. Les données sélectionnées doivent être de type numérique. Si le libellé des variables a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

Qualitatives : activez cette option si vous voulez inclure une ou plusieurs variables explicatives qualitatives dans le modèle. Sélectionnez alors la ou les variables correspondantes sur la feuille Excel. Les données sélectionnées peuvent être de tout type, mais les données numériques sont automatiquement considérées comme nominales. Si le libellé des variables a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

Modèle : choisissez le type de fonction à utiliser (voir la section description).

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des variables : activez cette option si la première ligne des données sélectionnées (variables dépendantes et explicatives, libellés des observations, poids) contient un libellé.

Libellés des observations : activez cette option si vous voulez utiliser des libellés d'observations pour l'affichage des résultats. Si l'option « Libellés des variables » est activée, la première cellule de la sélection doit comprendre un en-tête. Si vous n'activez pas cette option, des libellés seront automatiquement créés (Obs1, Obs2, …).

Poids des observations : ce champ est à remplir impérativement si l'option « somme de binaires » a été choisie. Sinon ce champ n'est pas actif. Si un en-tête de colonne a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

Poids dans la régression : activez cette option si vous voulez pondérer l'influence des observations pour l'ajustement du modèle. Si vous n'activez pas cette option, les poids seront tous considérés comme valant 1. Les poids doivent être impérativement supérieurs ou égaux à 0. Si un en-tête de colonne a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

Modalité de référence : dans le cas d'une régression logistique multinomiale, vous devez choisir ici qu'elle est la modalité de référence.

Onglet **Options** :

Sous-onglet **Générales** :

Algorithme

- **Itérations** : entrez le nombre maximal d'itérations pour l'algorithme de Newton-Raphson. Les calculs sont interrompus dès que le nombre maximal d'itérations est dépassé. Valeur par défaut : 100.
- **Convergence** : entrez la valeur seuil d'évolution maximale de log vraisemblance d'une itération à l'autre, qui une fois atteinte permet de considérer que l'algorithme a convergé. Valeur par défaut : 0,000001.
- **Temps maximum (s)** : entrez le temps maximum en secondes que vous être prêt(e) à consacrer à la recherche de l'optimum. Valeur par défaut : 180.
- **Pénalisation** : Vous pouvez choisir de ne pas pénaliser la vraisemblance (option par défaut) ou d'utiliser la pénalisation de **Firth** ou la pénalisation **L2** (voir la section

description). Dans le second cas, veuillez entrer la constante lambda de pénalisation (Valeur par défaut : 0.01).

Remarque : La méthode de First n'est pas disponible pour les variables dépendantes multinomiales et ordinales.

Interactions / Niveau : activez cette option pour inclure des interactions dans le modèle puis entrez le niveau maximum d'interaction (valeur comprise entre 1 et 4).

Tolérance : entrez la valeur de la tolérance seuil en deçà de laquelle une variable est automatiquement ignorée.

Intervalle de confiance (%) : entrez l'étendue en pourcentage de l'intervalle de confiance à utiliser pour les différents tests, et pour le calcul des intervalles de confiance autour des paramètres et des prédictions. Valeur par défaut : 95.

Intervalle de confiance LR : activez cette option pour calculer les intervalles de confiance LR.

Sous-onglet **Avancées**:

Sélection du modèle : activez cette option si vous souhaitez utiliser l'une des cinq méthodes de sélection proposées :

- **Meilleur modèle** : cette méthode permet de choisir le meilleur modèle parmi tous les modèles comprenant un nombre de variables variant de « Min variables » à « Max variables ». Par ailleurs le « critère » pour déterminer le meilleur modèle peut être choisi par l'utilisateur.
- **Critère** : veuillez choisir le critère parmi la liste suivante : Vraisemblance, LR (*likelihood ratio*), Score, Wald, AIC de Akaike, SBC de Schwarz.
- **Min variables** : entrez le nombre minimum de variables à prendre en compte dans le modèle.
- **Max variables** : entrez le nombre maximum de variables à prendre en compte dans le modèle.

Remarque : bien que grâce à un algorithme très performant XLSTAT réduise au maximum la quantité de calculs nécessaires, cette méthode peut entraîner des temps de calculs importants. Elle n'est pas disponible pour les modèles logit multinomial et ordinal.

- **Stepwise (Ascendante)** : le processus de sélection commence par l'ajout de la variable ayant la plus forte contribution au modèle. Si une seconde variable est telle que sa probabilité d'entrée est supérieure à la **valeur seuil pour entrer**, alors elle est ajoutée au modèle. A partir de l'ajout de la troisième variable, après chaque ajout, on évalue pour toutes les variables présentes dans le modèle quel serait l'impact de son retrait. Si la probabilité de la statistique calculée est supérieure à la **valeur seuil pour retirer**, la variable est retirée du modèle.
- **Stepwise (Descendante)** : cette méthode est similaire à la précédente, mais part d'un modèle complet.
- **Ascendante** : la procédure est identique à celle de la sélection progressive (*stepwise*), hormis le fait que les variables sont uniquement ajoutées et jamais retirées.

- **Descendante** : la procédure commence par l'ajout simultané de toutes les variables. Les variables sont ensuite retirées du modèle suivant la procédure utilisée pour la sélection progressive.

Correction du poids des classes : si les effectifs des différentes classes de la variable réponse ne sont pas homogènes, on risque de pénaliser dans le modèle les classes ayant un faible effectif. Afin de palier ce problème, XLSTAT propose deux options :

- **Automatique** : le redressement est automatique. Des poids artificiels sont affectés aux observations dans le but d'obtenir des classes dont la somme des poids est identique.
- **Poids correctifs** : vous pouvez sélectionner les poids à affecter à chacune des observations.

Contraintes : cette option n'est visible que si des variables explicatives qualitatives ont été sélectionnées. Des détails sur les différentes options sont disponibles dans la section description.

- **a1 = 0** : choisissez cette option pour que le paramètre de la première modalité de chaque facteur soit fixé à 0.
- **an = 0** : choisissez cette option pour que le paramètre de la dernière modalité de chaque facteur soit fixé à 0.
- **Somme (ai) = 0** : pour chaque facteur la somme des paramètres associés aux différentes modalités vaut 0.
- **Somme (ni.ai) = 0** : pour chaque facteur la somme des paramètres associés aux différentes modalités pondérés par la fréquence des modalités respectives vaut 0.

Effets imbriqués : activez cette option pour inclure un effet imbriqué dans le modèle.

Onglet **Validation** :

Validation : activez cette option si vous souhaitez utiliser une partie des données sélectionnées pour valider le modèle.

Jeu de validation : choisissez l'une des options pour définir le mode de sélection des observations utilisées pour la validation :

- **Aléatoire** : les observations sont sélectionnées de manière aléatoire. Le « Nombre d'observations » doit alors être saisi.
- **N dernières lignes** : les N dernières observations sont sélectionnées pour la validation. Le « Nombre d'observations » N doit alors être saisi.
- **N premières lignes** : les N premières observations sont sélectionnées pour la validation. Le « Nombre d'observations » N doit alors être saisi.
- **Variable de groupe** : si vous choisissez cette option, vous devez ensuite sélectionner une variable indicatrice composée de 0 pour les observations à utiliser pour le calcul du modèle, et de 1 pour les observations à utiliser pour la validation du modèle.

Onglet **Prédiction** :

Prédiction : activez cette option si vous souhaitez sélectionner des données à utiliser en mode prédiction. Si vous activez cette option, vous devez veiller à ce que les données de prédiction soient organisées comme les données d'estimation : mêmes variables, même ordre dans les sélections. En revanche vous ne devez pas sélectionner de libellés de variables : la première ligne des sélections décrites ci-dessous doit être une ligne de données.

Quantitatives : cette option est activée si votre modèle comprend des variables explicatives quantitavtives.

Qualitatives : cette option est activée si votre modèle comprend des variables explicatives qualitavtives.

Libellés des observations : activez cette option si vous voulez utiliser des libellés d'observations disponibles sur une feuille Excel pour l'affichage des résultats. La première ligne ne doit pas comprendre d'en-tête. Si vous n'activez pas cette option, des libellés seront automatiquement créés (PredObs1, PredObs2, …).

Libellés des variables : activez cette option si la première ligne des données sélectionnées pour les prédictions contient un libellé.

Onglet **Données manquantes** :

Ne pas accepter les données manquantes : activez cette option si vous voulez que les calculs soient interrompus et que vous soyez prévenu en cas de présence de données manquantes.

Supprimer les observations : activez cette option pour supprimer les observations comportant des données manquantes.

Estimer les données manquantes : activez cette option pour estimer les données manquantes avant le début des calculs.

- **Moyenne ou mode** : activez cette option pour estimer les données manquantes en utilisant la moyenne (variables quantitatives) ou le mode (variables qualitatives) pour les variables correspondantes.
- **Plus proche voisin** : activez cette option pour estimer les données manquantes d'une observation en recherchant le plus proche voisin de l'observation.

Onglet **Sorties**:

Statistiques descriptives : activez cette option pour afficher les statistiques descriptives pour les variables sélectionnées.

Corrélations : activez cette option pour afficher la matrice de corrélation des variables explicatives.

Statistiques de multicolinéarité : activez cette option pour afficher le tableau des statistiques de multicolinéarité.

Coefficients d'ajustement : activez cette option pour afficher le tableau des statistiques d'ajustement du modèle.

Analyse de type II : activez cette option pour afficher le tableau d'analyse de l'impact du retrait des variables sur le modèle (cette approche correspond à l'analyse de type II avec le modèle linéaire).

Test de Hosmer-Lemeshow : activez cette option pour afficher les résultats du test de Hosmer-Lemeshow.

Coefficients du modèle : activez cette option pour afficher le tableau des coefficients du modèle. Optionnellement les **intervalles de confiance** de type « *profile likelihood* » peuvent être calculés (voir description).

Equation : activez cette option pour afficher explicitement l'équation du modèle.

Coefficients normalisés : activez cette option pour afficher les paramètres normalisés du modèle (coefficients bêta).

Matrice de covariance : activez cette option pour afficher la matrice de covariance des coefficients du modèle.

Effets marginaux à la moyenne : activez cette option pour calculer les effets marginaux à la moyenne. Ces effets permettent de mesurer l'impact de chaque variable explicative lorsque toutes les autres sont fixées à leur moyenne.

Prédictions et résidus : activez cette option pour afficher les prédictions et les résidus pour l'ensemble des observations.

- **Modèle indépendant** : activez cette option pour afficher les résultats correspondant au modèle de base (dit indépendant) où la prédiction correspond simplement à la fréquence observée de chaque classe.
- **Intervalles de confiance** : activez cette option pour afficher les intervalles de confiance. Ceux-ci ne sont affichés que dans les cas où la variable réponse est binomiale.
- **Analyse de significativé** : activez cette option pour étudier si la probabilité associée à la modalité prédite est significativement différente de celle calculée pour d'autres modalités (voir la section description).

Tableau de classification : activez cette option pour afficher le tableau de classement a posteriori des observations sur la base d'un **point de séparation** à définir (valeur par défaut 0.5).

Analyse des probabilités : si une seule variable explicative a été sélectionnée, activez cette option pour que XLSTAT calcule la valeur de la variable explicative correspondant à divers niveaux de probabilité.

Comparaisons multiples : cette option n'est active que si des variables explicatives qualitatives ont été sélectionnées. Activez cette option pour afficher les résultats des tests de comparaison.

Onglet **Graphiques** :

Graphiques de régression : activez cette option pour afficher les graphiques de régression :

- **Coefficients normalisés** : activez cette option pour afficher sur un graphique les paramètres normalisés du modèle avec leur intervalle de confiance.
- **Prédictions** : activez cette option pour afficher la courbe de régression.
- **Intervalles de confiance** : activez cette option pour afficher les intervalles de confiance sur les graphiques (1) et (4).

Graphique de confusion : activez cette option pour afficher le graphique de confusion qui permet une visualisation synthétique du tableau de classification. Les effectifs peuvent être liés soit à la largeur, soit l'aire, des carrés représentés.

Résultats

XLSTAT propose un nombre important de tableaux et de graphiques afin de faciliter l'analyse et l'interprétation des résultats.

Statistiques descriptives : le tableau de statistiques descriptives présente pour toutes les variables sélectionnées des statistiques simples. Pour les variables quantitatives sont affichés le nombre d'observations, le nombre de données manquantes, le nombre de données non manquantes, la moyenne et l'écart-type (non biaisé). Pour les variables qualitatives, dont la variable dépendante, sont affichées les modalités, leurs effectifs et pourcentage respectifs.

Matrice de corrélation : dans ce tableau sont affichées les corrélations entre les variables explicatives. Il est à noter que si la variable dépendante est binaire, le coefficient de corrélation bisérielle est utilisé pour calculer la corrélation entre les variables explicatives quantitatives et la variable dépendante.

Synthèse de la sélection des variables : dans le cas où une méthode de sélection a été choisie, XLSTAT affiche la synthèse de la sélection. Dans le cas d'une sélection pas à pas, les statistiques correspondant aux différentes étapes sont affichées. Dans le cas d'une sélection du meilleur modèle pour un nombre de variables variant de p à q, le meilleur modèle pour chaque nombre de variable est affiché avec les statistiques correspondantes ; le meilleur modèle pour le critère choisi est alors affiché en gras.

Coefficients d'ajustement : dans ce tableau est affichée une série de statistiques pour le modèle indépendant (correspondant au cas où la combinaison linéaire des variables explicatives se réduit à une constante) et pour le modèle ajusté.

- **Observations** : le nombre total d'observations prises en compte (somme des poids des observations) ;
- **Somme des poids** : le nombre total d'observations prises en compte (somme des poids des observations multipliée par les poids dans la régression) ;

- **DDL** : degrés de liberté ;
- **-2 Log(Vrais.)** : le logarithme de la fonction de vraisemblance associée au modèle;
- **R² (McFadden) :** coefficient compris comme le R² entre 0 et 1 qui mesure le bon ajustement du modèle. Ce coefficient est égal à 1 moins le rapport de la vraisemblance du modèle ajusté sur la vraisemblance du modèle indépendant ;
- **R²(Cox et Snell) :** coefficient compris comme le R² entre 0 et 1 qui mesure le bon ajustement du modèle. Ce coefficient est égal à 1 moins le rapport de la vraisemblance du modèle ajusté sur la vraisemblance du modèle indépendant, le rapport étant porté à l'exposant 2/Sw, où Sw est la somme des poids ;
- **R²(Nagelkerke) :** coefficient compris comme le R² entre 0 et 1 qui mesure le bon ajustement du modèle. Ce coefficient est égal au rapport du R² de Cox et Snell, divisé par 1 moins la vraisemblance du modèle indépendant portée à l'exposant 2/Sw ;
- **AIC** : le critère d'information d'Akaike (Akaike's Information Criterion) ;
- **SBC** : le critère bayésien de Schwarz (Schwarz's Bayesian Criterion).

Test de l'hypothèse nulle H0 : **Y=p0** : l'hypothèse H0 correspond au modèle indépendant qui donne la probabilité $p0$ quelques soient les valeurs des variables explicatives ; on cherche à vérifier si le modèle ajusté est significativement plus performant que ce modèle. Trois tests sont proposés : le test du rapport des vraisemblance (-2 Log(Vrais.)), le test du Score, et le test test de Wald. Les trois statistiques suivent une loi du χ^2 dont les degrés de liberté sont indiqués.

Analyse de Type II : ce tableau n'a d'intérêt que s'il y a plus d'une variable explicative. On teste ici le modèle ajusté contre un modèle dont on aurait retiré la variable de la ligne du tableau en question. Si la probabilité $Pr > LR$ est inférieure à un seuil de signification que l'on se fixe (typiquement 0.05), alors la contribution de la variable à l'ajustement du modèle est significative. Sinon, elle peut être retirée du modèle.

Paramètres du modèle :

- **Cas binaire** : pour la constante du modèle et pour chaque variable explicative sont affichés l'estimation du paramètre, l'écart-type correspondant, le χ^2 de Wald, la p-value correspondante, ainsi que l'intervalle de confiance. Si l'option correspondante a été activée, les intervalles LR (*Likelihood Ratio*) sont aussi affichés. L'odds ratio et l'intervalle de confiance associé sont affichés dans la partie droite du tableau.
- **Cas Multinomial** : dans le cas multinomial, on obtient une série de coefficients pour chaque modalité active. On aura donc $(J-1)(q+1)$ lignes dans le tableau où J est le nombre de modalités de la variable dépendante et q est le nombre de variables explicatives. Ainsi, pour chaque variable et pour chaque modalité sont affichés l'estimation du paramètre, l'écart-type correspondant, le χ^2 de Wald, la p-value correspondante, l'intervalle de confiance, l'odds ratio et l'intervalle de confiance associé.

- **Cas ordinal** : dans le cas ordinal, on obtient une constante pour chaque modalité et une seule série de coefficients. On aura donc $(J-1)+q$ lignes dans le tableau où J est le nombre de modalités de la variable cible et q est le nombre de variables explicatives. Ainsi, pour chaque variable et pour chaque modalité sont affichés l'estimation du paramètre, l'écart-type correspondant, le χ^2 de Wald, la p-value correspondante et l'intervalle de confiance.

Les **équations du modèle** sont ensuite affichées pour faciliter la lecture ou la réutilisation du modèle.

Le tableau des **coefficients normalisés** (aussi appelés coefficients bêta) permet de comparer le poids relatif des variables. Plus la valeur absolue d'un coefficient est élevée, plus le poids de la variable correspondante est important. Lorsque l'intervalle de confiance autour des coefficients normalisés comprend la valeur 0 (cela est facilement visible sur le graphique des coefficients normalisés), le poids d'une variable dans le modèle n'est pas significatif.

Les **effets marginaux** au point correspondant aux moyennes des variables explicatives sont alors affichées. Les effets marginaux sont particulièrement intéressants lorsqu'ils sont comparés les uns aux autres. En les comparant, on peut mesurer l'impact relatif de chaque variable au point donné. L'impact peut être interprété comme l'influence d'une petite variation de chaque variable explicative, sur la variable dépendante. Un intervalle de confiance calculé à l'aide de la méthode Delta est affiché. XLSTAT fournit ces résultats pour les variables quantitatives et qualitatives, qu'il s'agisse de facteurs simples ou d'interactions. Pour les variables qualitatives, l'effet marginal indique l'impact d'un changement de modalité (de la première modalité à la modalité d'intérêt).

Dans le tableau des **prédictions et résidus** sont donnés pour chaque observation, son poids, la valeur de la variable explicative quantitative (s'il n'y en a qu'une), la valeur observée de la variable dépendante, la prédiction du modèle, les mêmes valeurs divisées par le poids dans le cas somme(binaires), les probabilités associées à chaque modalité. Dans le cas binaire, l'intervalle de confiance associé aux probabilités est affiché. Si la modalité prédite est différente de celle observée, il est indiqué si cela est significatif ou non. Il est également indiqué si la probabilité associée à la probabilité maximale est significativement différente des autres.

Le tableau des **diagnostics d'influence** permet de d'évaluer l'impact de chaque observation sur la qualité du modèle ou sur la valeur des coefficients du modèle.

Tableau de classification : activez cette option pour afficher le tableau permettant de visualiser le pourcentage d'observations bien classées pour chacune des deux catégories. Si un échantillon de validation a été extrait, ce tableau est aussi affiché pour les données de validation.

Courbe ROC : la courbe ROC permet d'évaluer la performance du modèle au travers de l'aire sous la courbe (AUC) et de comparer plusieurs modèles entre eux (voir la section description pour plus de détails). Elle n'est affichée que dans le cas binaire.

Comparaison des modalités des variables qualitatives : si une ou plusieurs variables qualitatives explicatives ont été sélectionnées, les résultats des tests d'égalité des paramètres pris deux à deux des différentes modalités des variables qualitatives sont affichés.

Le tableau d'**analyse des probabilités** n'est affiché que si une seule variable explicative quantitative a été sélectionnée et si l'on est dans le cas binomial. Il permet de visualiser à quel niveau de la variable explicative correspond une probabilité donnée.

Exemple

Un exemple de régression logistique et d'application du modèle logit multinomial sont disponibles sur le Centre d'aide XLSTAT à l'adresse

- Régression logistique : http://www.xlstat.com/demo-logf.htm

- Modèle logit multinomial : http://www.xlstat.com/demo-logmultf.htm

- Modèle logit ordinal : http://www.xlstat.com/demo-logordf.htm

Bibliographie

Agresti A. (2002). Categorical Data Analysis, 2-nd Edition. John Wiley and Sons, New York.

Finney D.J. (1971). Probit Analysis, 3rd Edition. Cambridge, London and New York.

Firth D (1993). Bias reduction of maximum likelihood estimates. *Biometrika*, **80**, 27-38.

Fomby T. B. and Pearce J. E. (1986). Standard errors in the multinomial logit model. *Communications in Statistics - Theory and Methods*, **15(8)**, 2555-2568.

Furnival G. M. and Wilson R.W. Jr. (1974). Regressions by leaps and bounds. *Technometrics*, **16** (4), 499-511.

Heinze G. and Schemper M. (2002). A solution to the problem of separation in logistic regression. *Statistics in Medicine*, **21**, 2409-2419.

Hosmer D.W. and Lemeshow S. (2000). Applied Logistic Regression, Second Edition. John Wiley and Sons, New York.

Lang J. B. (2014). The Pearson Score Statistic for Multinomial-Poisson Models. *Communications in Statistics - Theory and Methods*, **43(21)**, 4471-4491.

Lawless J.F. and Singhal K. (1978). Efficient screening of nonnormal regression Models. *Biometrics*, **34**, 318-327.

Lesaffre E. and Albert A. (1989). Multiple-group logistic regression diagnostics.*Journal of the Royal Statistical Society, Series C (Applied Statistics)*, **38(3)**, 425-440.

Pregibon D. (1981). Logistic regression diagnostics. *Annals of Statistics*, **9**, 705-724.

Sambamoorthi N., Ervin V.J. and Thomas G. (1994). Simultaneous prediction intervals for multinomial logistic regression models. *Communications in Statistics - Theory and Methods*, **23(3)**, 815-829.

Tallarida R.J. (2000). Drug Synergism & Dose-Effect Data Analysis. CRC/Chapman & Hall, Boca Raton.

Venzon, D. J. and Moolgavkar S. H. (1988). A method for computing profile likelihood Based confidence intervals. *Applied Statistics*, **37**, 87-94.

Régression log-linéaire

Cet outil permet d'ajuster un modèle de régression log-linéaire avec différentes lois de probabilité (Poisson, Gamma, Exponentielle).

Dans cette section :

Description

La régression log-linéaire permet de modéliser des données par une combinaison log-linéaire des paramètres du modèle et des variables explicatives (qui peuvent être qualitatives et/ou quantitatives). De plus, on suppose que les données (la réponse) sont distribuées soit selon une loi de Poisson, une loi Gamma ou une loi exponentielle.

Le modèle de régression log- linéaire

On note Y le vecteur de variables réponse et X la matrice des p covariables. La première colonne de la matrice X est composée de 1 et correspond à la constante ou l'ordonnée à l'origine (intercept). Le vecteur des paramètres est noté β. Le modèle s'écrit :

$$E(Y|X) = e^{\beta' X}$$

A partir de l'équation précédente, on obtient directement :

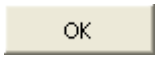
$$\log[E(Y|X)] = \beta' X$$

Inférence des paramètres du modèle

On suppose que les variables Y_i sont indépendantes avec X_i le vecteur de covariables associé, les paramètres du modèle peuvent être estimés par maximum de vraisemblance. Quel que soit la distribution choisie (Poisson, Gamma, Exponentielle), la fonction de vraisemblance est convexe et peut être maximisée par un algorithme de type Newton-Raphson.

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.

Onglet **Général**:

Variables dépendantes :

Variable(s) réponse : sélectionnez la ou les variables réponse que vous souhaitez modéliser. Si plusieurs variables sont sélectionnées, XLSTAT fera les calculs pour chacune des variables indépendamment. Si des en-têtes de colonnes ont été sélectionnés, veuillez vérifier que l'option « Libellés des variables » est activée ;

Variables explicatives :

Quantitatives : activez cette option si vous voulez inclure une ou plusieurs variables explicatives quantitatives dans le modèle. Sélectionnez alors la ou les variables correspondantes sur la feuille Excel. Les données sélectionnées doivent être de type numérique. Si le libellé des variables a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée ;

Qualitatives : activez cette option si vous voulez inclure une ou plusieurs variables explicatives qualitatives dans le modèle. Sélectionnez alors la ou les variables correspondantes sur la feuille Excel. Les données sélectionnées peuvent être de tout type, mais les données numériques sont automatiquement considérées comme nominales. Si le libellé des variables a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée ;

Offset : activez cette option si vous souhaitez inclure un offset. Cette option est disponible pour la distribution de Poisson.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des variables : activez cette option si la première ligne des données sélectionnées (variables dépendantes et explicatives, libellés des observations, poids) contient un libellé.

Libellés des observations : activez cette option si vous voulez utiliser des libellés d'observations pour l'affichage des résultats. Si l'option « Libellés des variables » est activée, la première cellule de la sélection doit comprendre un en-tête. Si vous n'activez pas cette option, des libellés seront automatiquement créés (Obs1, Obs2, …).

Poids dans la régression : activez cette option si vous voulez pondérer l'influence des observations pour l'ajustement du modèle. Si vous n'activez pas cette option, les poids seront tous considérés comme valant 1. Les poids doivent être impérativement supérieurs ou égaux à 0. Si un en-tête de colonne a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

Distributions : sélectionnez la distribution de probabilités que vous souhaitez utiliser pour modéliser vos données (Poisson, Gamma ou Exponentielle).

Onglet **Options**:

Tolérance : entrez la valeur de la tolérance seuil en deçà de laquelle une variable est automatiquement ignorée.

Intervalle de confiance (%) : entrez l'étendue en pourcentage de l'intervalle de confiance à utiliser pour les différents tests, et pour le calcul des intervalles de confiance autour des paramètres et des prédictions. Valeur par défaut : 95

Constante fixée : activez cette option pour fixer la valeur de la constante.

Interactions / Niveau : activez cette option pour inclure des interactions dans le modèle puis entrez le niveau maximum d'interaction (valeur comprise entre 1 et 4).

Conditions d'arrêt :

- **Itérations** : entrez le nombre maximal d'itérations pour l'algorithme de Newton-Raphson. Les calculs sont interrompus dès que le nombre maximal d'itérations est dépassé. Valeur par défaut : 100.
- **Convergence** : entrez la valeur seuil d'évolution maximale de log vraisemblance d'une itération à l'autre, qui une fois atteinte permet de considérer que l'algorithme a convergé. Valeur par défaut : 0.000001.

Contraintes : en présence de variables explicatives qualitatives, vous devez choisir ici qu'elle est la modalité de référence.

Sélection du modèle : activez cette option si vous souhaitez utiliser l'une des quatre méthodes de sélection proposées :

- **Stepwise (Ascendante)** : le processus de sélection commence par l'ajout de la variable ayant la plus forte contribution au modèle. Si une seconde variable est telle que sa probabilité d'entrée est supérieure à la **valeur seuil pour entrer**, alors elle est ajoutée au modèle. A partir de l'ajout de la troisième variable, après chaque ajout, on évalue pour toutes les variables présentes dans le modèle quel serait l'impact de son retrait. Si la probabilité de la statistique calculée est supérieure à la **valeur seuil pour retirer**, la variable est retirée du modèle ;
- **Stepwise (Descendante)** : cette méthode est similaire à la précédente, mais part d'un modèle complet.
- **Ascendante** : la procédure est identique à cette de la sélection progressive (*stepwise*), hormis le fait que les variables sont uniquement ajoutées et jamais retirées ;
- **Descendante** : la procédure commence par l'ajout simultané de toutes les variables. Les variables sont ensuite retirées du modèle suivant la procédure utilisée pour la sélection progressive ;
- **Critère** : veuillez choisir le critère parmi la liste suivante : LR (*Likelihood Ratio*), Wald ;

Onglet **Validation** :

Validation : activez cette option si vous souhaitez utiliser une partie des données sélectionnées pour valider le modèle.

Jeu de validation : choisissez l'une des options pour définir le mode de sélection des observations utilisées pour la validation :

- **Aléatoire** : les observations sont sélectionnées de manière aléatoire. Le « Nombre d'observations » doit alors être saisi.

- **N dernières lignes** : les N dernières observations sont sélectionnées pour la validation. Le « Nombre d'observations » N doit alors être saisi.
- **N premières lignes** : les N premières observations sont sélectionnées pour la validation. Le « Nombre d'observations » N doit alors être saisi.
- **Variable de groupe** : si vous choisissez cette option, vous devez ensuite sélectionner une variable indicatrice composée de 0 pour les observations à utiliser pour le calcul du modèle, et de 1 pour les observations à utiliser pour la validation du modèle.

Onglet **Prédiction** :

Prédiction : activez cette option si vous souhaitez sélectionner des données à utiliser en mode prédiction. Si vous activez cette option, vous devez veiller à ce que les données de prédiction soient organisées comme les données d'estimation : mêmes variables, même ordre dans les sélections. En revanche vous ne devez pas sélectionner de libellés de variables : la première ligne des sélections décrites ci-dessous doit être une ligne de données.

Quantitatives : activez cette option pour sélectionner la ou les variables quantitatives explicatives. La première ligne ne doit pas comprendre d'en- tête.

Qualitatives : activez cette option pour sélectionner la ou les variables qualitatives explicatives. La première ligne ne doit pas comprendre d'en-tête.

Libellés des observations : activez cette option si vous voulez utiliser des libellés d'observations disponibles sur une feuille Excel pour l'affichage des résultats. La première ligne ne doit pas comprendre d'en-tête. Si vous n'activez pas cette option, des libellés seront automatiquement créés (PredObs1, PredObs2, …).

Libellés des variables : activez cette option si la première ligne des données sélectionnées (variables explicatives, libellés des observations) contient un libellé.

Onglet **Données manquantes** :

Supprimer les observations : activez cette option pour supprimer les observations comportant des données manquantes.

Estimer les données manquantes : activez cette option pour estimer les données manquantes avant le début des calculs.

- **Moyenne ou mode** : activez cette option pour estimer les données manquantes en utilisant la moyenne (variables quantitatives) ou le mode (variables qualitatives) pour les variables correspondantes.
- **Plus proche voisin** : activez cette option pour estimer les données manquantes d'une observation en recherchant le plus proche voisin de l'observation.

Onglet **Sorties**:

Statistiques descriptives : activez cette option pour afficher les statistiques descriptives pour les variables sélectionnées.

Corrélations : activez cette option pour afficher la matrice de corrélation des variables explicatives.

Coefficients d'ajustement : activez cette option pour afficher le tableau des statistiques d'ajustement du modèle.

Analyse de type III : activez cette option pour afficher le tableau d'analyse de la variable de type III.

Coefficients du modèle : activez cette option pour afficher le tableau des coefficients du modèle.

Equation : activez cette option pour afficher explicitement l'équation du modèle.

Prédictions et résidus : activez cette option pour afficher les prédictions et les résidus pour l'ensemble des observations.

Test de surdispersion : activez cette option pour tester la surdispersion dans le cas d'une régression de Poisson.

Onglet **Graphiques** :

Graphiques de régression : activez cette option pour afficher les graphiques de régression.

- **Intervalles de confiance** : activez cette option pour afficher les intervalles de confiance.

Graphique des prédictions : activez cette option pour afficher les courbes de prédictions.

- **Intervalles de confiance** : activez cette option pour afficher les intervalles de confiance.

Résultats

XLSTAT propose un nombre important de tableaux et de graphiques afin de faciliter l'analyse et l'interprétation des résultats.

Statistiques descriptives : le tableau de statistiques descriptives présente pour toutes les variables sélectionnées des statistiques simples. Pour les variables quantitatives Sont affichés le nombre d'observations, le nombre de données manquantes, le nombre de données non manquantes, la moyenne, et l'écart-type (non biaisé). Pour les variables qualitatives, dont la variable dépendante, sont affichées les modalités leurs effectifs et pourcentage respectifs.

Matrice de corrélation : dans ce tableau sont affichées les corrélations entre les variables explicatives.

Correspondance entre les modalités de la variable réponse et les probabilités : ce tableau permet de visualiser à quelles modalités de la variable dépendante ont été affectées les probabilités 0 et 1. Il est affiché dans le cas de variables dépendantes binaires.

Synthèse de la sélection des variables : dans le cas où une méthode de sélection a été choisie, XLSTAT affiche la synthèse de la sélection. Dans le cas d'une sélection pas à pas, les statistiques correspondant aux différentes étapes sont affichées. Dans le cas d'une sélection du meilleur modèle pour un nombre de variables variant de p à q, le meilleur modèle pour chaque nombre de variable est affiché avec les statistiques correspondantes ; le meilleur modèle pour le critère choisi est alors affiché en gras.

Coefficients d'ajustement : dans ce tableau est affichée une série de statistiques pour le modèle indépendant (correspondant au cas où la combinaison linéaire des variables explicatives se réduit à une constante) et pour le modèle ajusté.

- **Observations** : le nombre total d'observations prises en compte (somme des poids des observations) ;
- **Somme des poids** : le nombre total d'observations prises en compte (somme des poids des observations multipliés par les poids dans la régression) ;
- **DDL** : degrés de liberté ;
- **-2 Log(Vrais.)** : le logarithme de la fonction de vraisemblance associée au modèle;
- **R² (McFadden) :** coefficient compris comme le R² entre 0 et 1 qui mesure le bon ajustement du modèle. Ce coefficient est égal à 1 moins le rapport de la vraisemblance du modèle ajusté sur la vraisemblance du modèle indépendant ;
- **R²(Cox et Snell) :** coefficient compris comme le R² entre 0 et 1 qui mesure le bon ajustement du modèle. Ce coefficient est égal à 1 moins le rapport de la vraisemblance du modèle ajusté sur la vraisemblance du modèle indépendant, le rapport étant porté à l'exposant 2/Sw, où Sw est la somme des poids ;
- **R²(Nagelkerke) :** coefficient compris comme le R² entre 0 et 1 qui mesure le bon ajustement du modèle. Ce coefficient est égal au rapport du R² de Cox et Snell, divisé par 1 moins le la vraisemblance du modèle indépendant portée à l'exposant 2/Sw ;
- **Déviance :** critère de déviance pour le modèle indépendant et le modèle ajusté ;
- **Chi-2 de Pearson :** valeur du Chi-2 de Pearson pour le modèle indépendant et le modèle ajusté ;
- **AIC** : le critère d'information d'Akaike (Akaike's Information Criterion) ;
- **SBC** : le critère bayésien de Schwarz (Schwarz's Bayesian Criterion).

Test de nullité des coefficients : on cherche à vérifier si le modèle ajusté est significativement plus performant que le modèle indépendant. Trois tests sont proposés : le test du rapport des vraisemblances (-2 Log(Vrais.)), le test du Score, et le test de Wald. Les trois statistiques suivent une loi du Khi² dont les degrés de liberté sont indiqués.

Analyse de Type III : ce tableau n'a d'intérêt que s'il y a plus d'une variable explicative. On teste ici le modèle ajusté contre un test dont on aurait retiré la variable de la ligne du tableau en

question. Si la probabilité Pr > LR est inférieur à un seul de signification que l'on se fixe (typiquement 0.05), alors la contribution de la variable à l'ajustement du modèle est significative. Sinon, elle peut être retirée du modèle.

Paramètres du modèle : pour la constante du modèle et pour chaque variable sont affichés l'estimation du paramètre, l'écart-type correspondant, le Khi² de Wald, la p-value correspondante, ainsi que l'intervalle de confiance.

Exemple

Un exemple d'application de la régression log-linéaire est disponible sur le Centre d'aide XLSTAT :

http://www.xlstat.com/demo-LogLinRegf.htm

Bibliographie

Ter Berg P. (1980). On the loglinear poisson and Gamma model. *Astin Bulletin*, **11**, 35-40.

Régression Quantile

Utilisez la régression quantile pour modéliser une variable dépendante quantitative en fonction de variables explicatives quantitatives ou qualitatives. L'usage de la régression quantile permet une analyse plus fine que celle fournie par la régression classique ou par l'ANCOVA en étendant l'estimation de la moyenne conditionnelle à celle des quantiles conditionnels.

Dans cette section :

Description

La régression quantile est une méthode de plus en plus utilisée car elle offre la possibilité de se placer dans un cadre d'étude au plus proche de la réalité.

En effet, de par sa nature, elle permet de s'affranchir de la rigidité contraignante de l'hypothèse de normalité intrinsèque au modèle de régression classique, offrant ainsi la liberté de travailler avec un large spectre de distributions.

En outre, comme mentionné dans l'introduction, la régression quantile permet à l'utilisateur d'ajuster au mieux son analyse à la problématique considérée, via l'estimation des quantiles conditionnels les plus pertinents.

De ce fait, on pourra profiter et tirer parti de cet outil dans des domaines d'application très variés. On notera tout de même, en s'appuyant sur le nombre de publications, que les études basées sur la régression quantile touchent plus particulièrement les sciences du Vivant/Santé, de l'Economie/Finance, de l'Environnement/Ecologie, sans oublier les Sciences Sociales, Comportementales ou Cognitives.

D'un point de vue historique, la régression quantile a été introduite en 1978 par Koenker et Basset. Depuis, l'engouement pour cette technique n'a cessé de croître, en attestent les très nombreux développements sur le sujet dont les principales références figurent dans la bibliographie.

Modèle

Comme dans le cadre de l'ANCOVA, la variable dépendante Y est quantitative tandis que l'ensemble des régresseurs X peut aussi bien être constitué de variables quantitatives que de facteurs (variables quantitatives + interactions entre les variables, quelles qu'elles soient).

Cependant, il est important d'avoir à l'esprit que dans cette méthode, aucune hypothèse sur la distribution des erreurs n'est requise.

Formalisation mathématique du problème

Le quantile d'ordre α , $\alpha \in [0,1]$, est défini comme étant la valeur y tel que : $P(Y = y) = \alpha$. En introduisant la fonction de répartition F, on définit la fonction quantile Q comme étant son inverse :

$$Q(\alpha) = F^{-1}(\alpha) = \inf\{y : F(y) > \alpha\}$$

La moyenne μ de la v.a. Y peut être vue comme étant la valeur telle que :

$$\mu = \operatorname{argmin}\{c : E[(Y - c)^2]\}$$

(1)

minimisant ainsi la somme des carrés des déviations.

De la même manière, une caractérisation du quantile d'ordre α , q_α, est possible en observant que :

$$q_\alpha = \operatorname{argmin}\{c : E[\rho_\alpha(Y - c)]\}$$

(2)

où ρ_α désigne la fonction suivante :

$$\begin{aligned}\rho_\alpha &= [\alpha - I_{\{y<0\}}]y \\ &= [(1-\alpha)I_{\{y<0\}} + \alpha I_{\{y>0\}}]|y|\end{aligned}$$

Ainsi, q_α minimise une somme pondérée de déviations prises en valeur absolue.

Si maintenant on se replace dans le contexte où Y est une variable réponse et X un ensemble de variables explicatives, dans le cadre linéaire, le problème de minimisation (1) devient :

$$\hat{\beta}_\alpha = \operatorname{argmin}\left\{\beta_\alpha : E\left[\left(Y - x_i^T\beta_\alpha\right)^2\right]\right\}$$

De la même manière, on peut récrire (2) en :

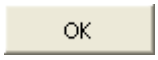
$$\hat{\beta}_\alpha = \operatorname{argmin}\left\{\beta_\alpha : E\left[\rho_\alpha\left(Y - X\beta_\alpha\right)\right]\right\}$$

où les paramètres et les estimateurs associés dépendent de l'ordre α considéré.

Ainsi, alors que le problème classique de minimisation au sens des moindres carrés nous amène à considérer la moyenne conditionnelle de la variable dépendante, la formulation du problème de régression quantile va nous conduire à estimer des quantiles conditionnels.

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.

Onglet **Général** :

Y / Variables dépendantes :

Quantitatives : sélectionnez la ou les variables réponse que vous souhaitez modéliser. Si plusieurs variables sont sélectionnées, XLSTAT fera les calculs pour chacune des variables indépendamment. Si des en-têtes de colonnes ont été sélectionnés, veuillez vérifier que l'option « Libellés des variables » est activée.

X / Variables explicatives :

Quantitatives : sélectionnez la ou les variables qualitatives explicatives sur la feuille Excel. Les données sélectionnées doivent être de type numérique. Si le libellé des variables a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

Qualitatives : activez cette option si vous voulez inclure une ou plusieurs variables explicatives qualitatives dans le modèle. Sélectionnez alors la ou les variables correspondantes sur la feuille

Excel. Les données sélectionnées peuvent être de tout type, mais les données numériques sont automatiquement considérées comme nominales. Si le libellé des variables a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des variables : activez cette option si la première ligne des données sélectionnées (variables dépendantes et explicatives, libellés des observations, poids) contient un libellé.

Libellés des observations : activez cette option si vous voulez utiliser des libellés d'observations pour l'affichage des résultats. Si l'option « Libellés des variables » est activée, la première cellule de la sélection doit comprendre un en-tête. Si vous n'activez pas cette option, des libellés seront automatiquement créés (Obs1, Obs2, …).

Poids des observations : activez cette option si vous voulez pondérer les observations. Si vous n'activez pas cette option, les poids seront tous considérés comme valant 1. Les poids doivent être impérativement supérieurs ou égaux à 0. Un poids de 2 est équivalent à répéter deux fois la même observation. Si un en-tête de colonne a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

Quantile(s) :

Sélection : activez cette option si vous souhaitez étudier une sélection de quantiles. Sélectionnez alors les cellules contenant l'ordre α des quantiles d'intérêt sur la feuille de travail Excel.

Processus : activez cette option si vous souhaitez mettre en œuvre le processus quantile complet. Le nombre de quantiles effectivement calculés est déterminé par une heuristique dépendant (de façon croissante) à la fois du nombre d'observations et du nombre de régresseurs. En fonction de ce nombre, les ordres des quantiles calculés sont répartis de façon uniforme sur $[0, 1]$.

Onglet **Options** :

Algorithme : 3 algorithmes différents sont proposés pour le calcul des coefficients de la régression quantile !

- **Simplexe** : mise en œuvre de l'algorithme de Barrodale et Roberts basé sur des méthodes de simplexe. Son usage est recommandé lorsque le nombre d'observations est inférieur à 5000 et le nombre de variables est inférieur à 50.
- **Point intérieur** : mise en œuvre de l'algorithme prédicteur-correcteur de Mehrotra basé sur des méthodes de point intérieur.
- **Fonction régularisante :** mise en œuvre de l'algorithme de Clark et Osborne basé sur l'approximation de la fonction objectif par une fonction plus régulière dont la minimisation donne, asymptotiquement, les mêmes résultats que celle de la fonction d'origine. Cet algorithme est très compétitif, surtout pour les jeux de données où $\frac{p}{n} > 0,05$.

Conditions d'arrêt : l'algorithme sélectionné s'arrête quand le premier des 3 évènements suivants se réalise :

- Fin de l'algorithme OU
- Nombre d'itérations maximum que vous pouvez définir dans **Itérations** atteint OU
- Seuil de convergence que vous pouvez définir dans **Convergence** atteint.

Intervalle de confiance (%) : entrez l'étendue en pourcentage de l'intervalle de confiance à utiliser pour les différents tests, et pour le calcul des intervalles de confiance autour des paramètres et des prédictions. Valeur par défaut : 95.

Interactions / Niveau : activez cette option pour inclure des interactions dans le modèle puis entrez le niveau maximum d'interaction (valeur comprise entre 1 et 4).

Contraintes : au cours des calculs, chaque facteur est décomposé en une sous-matrice contenant autant de colonnes qu'il y a de modalités dans le facteur. Typiquement, il s'agit d'un tableau disjonctif complet. Cette décomposition pose néanmoins un problème : s'il y a g modalités, le rang de cette sous-matrice n'est pas g mais $g - 1$. Cela entraîne la nécessité de supprimer l'une des colonnes de la sous-matrice, et éventuellement de transformer les autres colonnes. Deux stratégies sont alors possibles, dans le cadre de la régression quantile, en fonction de l'interprétation que l'on veut ensuite faire :

1) **a1=0** : le paramètre correspondant à la première modalité est nul. Ce choix permet d'imposer que l'effet de la première modalité correspond à un standard. Dans ce cas, la constante du modèle est égale à la moyenne de la variable dépendante pour le groupe 1.

2) **an=0** : le paramètre correspondant à la dernière modalité est nul. Ce choix permet d'imposer que l'effet de la dernière modalité correspond à un standard. Dans ce cas, la constante du

modèle est égale à la moyenne de la variable dépendante pour le groupe g.

Type d'erreur a priori : cochez cette option si vous avez une idée sur l'erreur a priori. Sélectionnez ensuite son type : erreur homogène (i.i.d.), hétérogène (i.n.i.d.) ou dépendante (n.i.i.d.) (par ex. erreurs autocorrélées). Cette sélection entrera ensuite en compte dans le calcul de la matrice de covariance des coefficients de la régression quantile, des intervalles de confiance,… si vous avez conjointement coché l'option **Distribution asymptotique** dans l'onglet **Sorties**.

Onglet **Validation** :

Validation : activez cette option si vous souhaitez utiliser une partie des données sélectionnées pour valider le modèle.

Jeu de validation : choisissez l'une des options pour définir le mode de sélection des observations utilisées pour la validation :

- **Aléatoire** : les observations sont sélectionnées de manière aléatoire. Le « Nombre d'observations » doit alors être saisi.
- **N dernières lignes** : les N dernières observations sont sélectionnées pour la validation. Le « Nombre d'observations » N doit alors être saisi.
- **N premières lignes** : les N premières observations sont sélectionnées pour la validation. Le « Nombre d'observations » N doit alors être saisi.
- **Variable de groupe** : si vous choisissez cette option, vous devez ensuite sélectionner une variable indicatrice composée de 0 pour les observations à utiliser pour le calcul du modèle, et de 1 pour les observations à utiliser pour la validation du modèle.

Onglet **Prédiction** :

Prédiction : activez cette option si vous souhaitez sélectionner des données à utiliser en mode prédiction. Si vous activez cette option, vous devez veiller à ce que les données de prédiction soient organisées comme les données d'estimation : mêmes variables, même ordre dans les sélections. En revanche vous ne devez pas sélectionner de libellés de variables : la première ligne des sélections décrites ci-dessous doit être une ligne de données.

Quantitatives : activez cette option pour sélectionner la ou les variables quantitatives explicatives. La première ligne ne doit pas comprendre d'en- tête.

Qualitatives : activez cette option pour sélectionner la ou les variables qualitatives explicatives. La première ligne ne doit pas comprendre d'en-tête.

Libellés des observations : activez cette option si vous voulez utiliser des libellés d'observations disponibles sur une feuille Excel pour l'affichage des résultats. La première ligne ne doit pas comprendre d'en-tête. Si vous n'activez pas cette option, des libellés seront automatiquement créés (PredObs1, PredObs2, …).

Libellés des variables : activez cette option si la première ligne des données sélectionnées (variables explicatives, libellés des observations) contient un libellé.

Onglet **Données manquantes** :

Supprimer les observations : activez cette option pour supprimer les observations comportant des données manquantes.

Estimer les données manquantes : activez cette option pour estimer les données manquantes avant le début des calculs.

- **Moyenne ou mode** : activez cette option pour estimer les données manquantes en utilisant la moyenne (variables quantitatives) ou le mode (variables qualitatives) pour les variables correspondantes.
- **Plus proche voisin** : activez cette option pour estimer les données manquantes d'une observation en recherchant le plus proche voisin de l'observation.

Onglet **Sorties** :

Statistiques descriptives : activez cette option pour afficher les statistiques descriptives. Les tableaux de statistiques descriptives présentent pour toutes les variables sélectionnées des statistiques simples. Pour les variables dépendantes (en bleu) et les variables explicatives quantitatives, sont affichés le nombre d'observations, le nombre de données manquantes, le nombre de données non manquantes, la moyenne, et l'écart-type (non biaisé). Pour les variables explicatives qualitatives sont affichés le nom des différentes modalités ainsi que leur fréquence respective.

Corrélations quantiles : activez cette option pour afficher la matrice des corrélations quantiles pour les variables quantitatives (dépendantes et explicatives).

Matrice de covariance : activez cette option pour afficher la matrice de covariance des coefficients.

Coefficients d'ajustement : activez cette option pour afficher les statistiques relatives à l'ajustement du modèle de régression

Tests de significativité du modèle : activez cette option pour effectuer un des tests disponibles pour évaluer la significativité du modèle. Plus précisément, on teste l'hypothèse du modèle complet contre l'hypothèse du modèle uniquement formé de la constante. 3 tests sont disponibles :

- **RV** : activez cette option pour effectuer un test de Rapport de Vraisemblance,
- **ML** : activez cette option pour effectuer un test des Multiplicateurs de Lagrange,
- **Wald** : activez cette option pour effectuer un test de Wald.

Equation du modèle : activez cette option pour afficher l'équation du modèle.

Prédictions et résidus : activez cette option pour afficher les prédictions et les résidus pour l'ensemble des observations. Dans le tableau des **prédictions et résidus** sont donnés pour chaque observation, son poids, la valeur de la variable explicative qualitative s'il n'y en a qu'une, la valeur observée de la variable dépendante, la prédiction du modèle, les résidus.

Calculs basés sur :

- **Distribution asymptotique** : activez cette option pour que le calcul de la matrice de covariance empirique des coefficients de la régression quantile ainsi que des intervalles de confiance soit effectué en considérant la distribution asymptotique théorique des coefficients. Ce calcul prendra en considération le **type d'erreur a priori** de l'onglet **Options** si vous l'avez renseigné.
- **Ré-échantillonnage (Bootstrap)** : activez cette option pour que le calcul de la matrice de covariance empirique des coefficients de la régression quantile ainsi que des intervalles de confiance soit effectué en introduisant de la variabilité via un ré-échantillonnage avec remise (Bootstrap). Si cette option est activée, rentrez un nombre entier dans **B =** pour choisir le nombre d'échantillons sur lequel sera basé le calcul des estimations.
- **Largeur de bande de Hall et Sheather** : activez cette option pour que le calcul de la matrice de covariance empirique des coefficients de la régression quantile ainsi que des intervalles de confiance soit effectué en utilisant la largeur de bande de Hall et Sheather $\left(O(n^{-1/3})\right)$.
- **Largeur de bande de Bofinger** : activez cette option pour que le calcul de la matrice de covariance empirique des coefficients de la régression quantile ainsi que des intervalles de confiance soit effectué en utilisant la largeur de bande de Bofinger $\left(O(n^{-1/5})\right)$.

Onglet **Graphiques** :

Graphiques de régression : activez cette option pour afficher les graphiques de régression :

- **Prédictions et résidus** : activez cette option pour afficher les graphiques suivants :
- Variable explicative versus résidus normalisés : ce graphique n'est affiché que s'il n'y a qu'une seule variable explicative et si cette variable est quantitative.
- Variable dépendante versus résidus normalisés.
- Prédictions pour la variable dépendante versus variable dépendante.

Résultats

Si le processus quantile a été sélectionné dans l'onglet Général, un tableau général donnant la valeur des coefficients associés à chaque q-quantile est affiché.

Des graphiques représentant l'évolution de ces coefficients en fonction de la valeur de a sont alors tracés pour une meilleure visualisation des résultats.

Dans le cas où une sélection de quantiles a été choisie dans l'onglet Général, on dispose des résultats suivants :

Statistiques descriptives : les tableaux de statistiques descriptives présentent pour toutes les variables sélectionnées des statistiques simples. Pour les variables dépendantes (en bleu) et les variables explicatives quantitatives, sont affichés le nombre d'observations, le nombre de données manquantes, le nombre de données non manquantes, la moyenne, et l'écart-type (non biaisé). Pour les variables explicatives qualitatives sont affichés le nom des différentes modalités ainsi que leur fréquence respective.

Ensuite, pour chaque quantile, nous avons :

Coefficients d'ajustement : dans ce tableau sont affichées les statistiques relatives à l'ajustement du modèle de régression :

- Observations : le nombre d'observations prises en compte dans les calculs. Dans les formules présentées ci-dessous, n désigne le nombre d'observations.
- Somme des poids : la somme des poids des observations prises en compte dans les calculs. Dans les formules présentées ci-dessous W désigne la somme des poids.
- DDL : le nombre de degrés de liberté du modèle (correspondant à la partie erreurs).
- R^2_α : le coefficient de détermination du modèle. Ce coefficient, dont la valeur est comprise entre 0 et 1, n'est affiché que si la constante du modèle n'est pas fixée par l'utilisateur. Sa valeur est définie par :

$$\begin{aligned} R^2_\alpha &= 1 - \frac{SAR_\alpha}{SAT_\alpha} \\ &= 1 - \frac{\sum_{i=1}^{n} \rho_\alpha(y_i - x_i^T \hat{\beta}_\alpha)}{\sum_{i=1}^{n} \rho_\alpha(y_i - \hat{\beta}_{0,\alpha})} \end{aligned}$$

où $\hat{\beta}_{0,\alpha}$ est l'estimation de la constante du modèle (ordonnée à l'origine) pour le quantile d'ordre α .

SAR_α correspondant à la Somme en valeur Absolue Résiduelle et

SAT_α correspondant à la Somme en valeur Absolue Totale.

Le ${R^2}_\alpha$ s'interprète comme la proportion de variabilité de la variable dépendante expliquée par le modèle. Plus le ${R^2}_\alpha$ est proche de 1, meilleur est le modèle. L'inconvénient du ${R^2}_\alpha$ est qu'il ne prend pas en compte le nombre de variables utilisées pour ajuster le modèle.

- ${R^2}_\alpha$ ajusté : le coefficient de détermination ajusté du modèle. Le ${R^2}_\alpha$ ajusté peut être négatif si le ${R^2}_\alpha$ est voisin de zéro. Ce coefficient n'est affiché que si la constante du modèle n'est pas fixée par l'utilisateur. Sa valeur est définie par :

$$ {R^2}_\alpha \text{ajusté} = 1 - (1 - R^2_\alpha)\frac{W-1}{W-p-1} $$

- Le R^2_α ajusté est une correction du R^2_α qui permet de prendre en compte le nombre de variables utilisées dans le modèle.

- MAR_α: la Moyenne de la valeur Absolue des Résidus est définie par :

$$ MAR_\alpha = \frac{1}{W-p} SAR_\alpha $$

- $RMAR_\alpha$: la racine de la moyenne de la valeur absolue des résidus $(RMAR)$ est la racine carrée de la MAR_α.

- $MAPE_\alpha$: la Mean Absolute Percentage Error est calculée comme suit :

$$ MAPE_\alpha = \frac{100}{W} \sum_{i=1}^{n} w_i \left| \frac{y_i - \hat{y}_{i,\alpha}}{y_i} \right| $$

- Cp_α: le coefficient Cp_α de Mallows est défini par :

$$ Cp_\alpha = \frac{SAR_\alpha}{\hat{\sigma}_\alpha} + 2p - W $$

où SAR_α est la somme en valeur absolue résiduelle pour le modèle avec p variables explicatives, et où $\hat{\sigma}_\alpha$ correspond à l'estimateur de la variance des résidus pour le modèle comprenant toutes les variables explicatives. Plus le coefficient Cp_α est proche de p et moins le modèle est biaisé.

- AIC_α: le critère d'information d'Akaike (Akaike's Information Criterion) est défini par

$$ AIC_\alpha = W \ln\left(\frac{SAR_\alpha}{W}\right) + 2p $$

Ce critère proposé par Akaike (1973) dérive de la théorie de l'information, et s'appuie sur la mesure de Kullback et Leibler (1951). C'est un critère de sélection de modèles qui pénalise les modèles pour lesquels l'ajout de nouvelles variables explicatives n'apporte pas suffisamment d'information au modèle, l'information étant mesurée au travers de la SAR_α. On cherche à minimiser le critère AIC.

- SBC_α: le critère bayésien de Schwarz (Schwarz's Bayesian Criterion) est défini par

$$ SBC_\alpha = W \ln\left(\frac{SAR_\alpha}{W}\right) + \ln(W)p $$

Ce critère proposé par Schwarz (1978) est proche du critère AIC, et comme ce dernier on cherche à le minimiser.

- PC_α: le critère de prédiction d'Amemiya (Amemiya's Prediction Criterion) est défini par

$$PC_\alpha = \frac{(1 - R^2_\alpha)(W + p)}{W - p}$$

Ce critère proposé par Amemiya (1980) permet, comme le ${R^2}_\alpha$ ajusté, de tenir compte de la parcimonie du modèle.

Le tableau des **paramètres du modèle** permet de récapituler l'estimation des paramètres du modèle, l'erreur standard correspondante ainsi que les bornes de l'intervalle de confiance.

Le tableau de **significativité du modèle** permet d'évaluer le pouvoir explicatif des variables explicatives. Le pouvoir explicatif est évalué en comparant l'ajustement du modèle final avec l'ajustement du modèle rudimentaire composé d'une constante égale au quantile de la variable dépendante.

Dans le tableau des prédictions et résidus sont donnés pour chaque observation, son poids, la valeur de la variable explicative qualitative s'il n'y en a qu'une, la valeur observée de la variable dépendante, la prédiction du modèle, les résidus. Si des données de validation ont été sélectionnées, elles sont affichées en fin de tableau.

Les graphiques qui suivent permettent de visualiser les résultats mentionnés ci-dessus. S'il n'y a qu'une seule variable explicative dans le modèle, le premier graphique affiché permet de visualiser les valeurs observées, la droite de régression et les deux types d'intervalles de confiance autour des prévisions. Le second graphique permet, quant à lui, de visualiser les résidus normalisés en fonction de la variable explicative. En principe, les résidus doivent être distribués de manière aléatoire autour de l'axe des abscisses. L'observation d'une tendance ou d'une forme révèlerait un problème au niveau du modèle.

Exemple

Un exemple de régression quantile simple est disponible sur le Centre d'aide XLSTAT :

http://www.xlstat.com/demo-quantileregf.htm

Bibliographie

Barrodale I. and Roberts F.D.K. (1974). An improved algorithm for discrete L1 linear approximation. *SIAM Journal on Numerical Analysis* **10** , 839-848.

Chen C. (2007). A finite smoothing algorithm for quantile regression. *Journal of Computational and Graphical Statistics*, **16** (1), 136-164.

Clark D.I. and Osborne, M.R. (1986). Finite algorithms for Huber's M-estimator. *SIAM J. on Scientific and Statistical Computing*, **7**, 72-85.

Davino C., Furno M. and Vistocco D. (2013). Quantile Regression : Theory and Applications. John Wiley & Sons.

Koenker R. (2005). Quantile Regression. Cambridge University Press.

Koenker R. and D'Orey V. (1987). Algorithm AS 229: computing regression quantiles. *Journal of the Royal Statistical Society* : Series C (Applied Statistics) 36(3), 383–393.

Koenker R. and Machado J.A.F. (1999). Goodness of Fit and Related Inference Processes for Quantile Regression.*Journal of the American Statistical Association*. Vol. 94, n°448,1296-1310.

Mehrotra S. (1992). On the implementation of a primal–dual interior point method. *SIAM Journal on Optimization 2* (4): 575-60.

Splines cubiques

Cet outil permet d'ajuster une spline cubique à partir de nœuds définis par l'utilisateur.

Dans cette section :

Description

Une spline cubique est une fonction définie par morceaux par des polynômes de degré 3. Les splines cubiques sont très utilisées dans des problèmes d'interpolation où elles sont préférées aux méthodes classiques d'interpolation polynômiale. En effet, elles permettent de réaliser un compromis entre la régularité de la courbe et le degré des polynômes utilisés.

Spline cubique

Une spline cubique S est une fonction par morceaux définie sur un intervalle $[a, b]$ divisé en K intervalles $[x_{i-1}, x_i]$ tels que :

$$a = x_0 < x_1 < \ldots < x_{K-1} < x_K = b$$

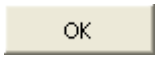
On note P_i le polynôme de degré 3 défini sur l'intervalle $[x_{i-1}, x_i]$. La spline S s'écrit alors sous la forme :

$$\begin{cases} S(t) = P_1(t) \text{ si } t \in [x_0, x_1] \\ \vdots \\ S(t) = P_K(t) \text{ si } t \in [x_{K-1}, x_K] \end{cases}$$

Le calcul des coefficients de la spline cubique repose sur le calcul des dérivées des différents polynômes (le détail de l'algorithme utilisé peut être trouvé dans Guillod, 2008).

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondants aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas (mode colonnes), XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite (mode lignes), XLSTAT considère que les variables sont en lignes et les observations en colonnes.

Onglet **Général**:

Y : sélectionnez les données qui correspondent aux y(i) dans l'équation de régression. Si un libellé est présent en première position, veillez à ce que l'option « Libellé des variables » soit bien activée.

X : sélectionnez les données qui correspondent aux x(i) dans l'équation de régression. Si un libellé est présent en première position, veillez à ce que l'option « Libellé des variables » soit bien activée.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des variables : activez cette option si la première ligne des données sélectionnées (variables dépendantes et explicatives, libellés des observations, poids) contient un libellé.

Libellés des observations : activez cette option si vous voulez utiliser des libellés d'observations pour l'affichage des résultats. Si l'option « Libellés des variables » est activée, la première cellule de la sélection doit comprendre un en-tête. Si vous n'activez pas cette option, des libellés seront automatiquement créés (Obs1, Obs2, …).

Onglet **Options**:

- **Données en tant que nœuds** : activez cette option pour que les données sélectionnées soient utilisées comme les nœuds de la spline.
- **Nombre de nœuds** : activez cette option pour sélectionner le nombre de nœuds de la spline. Ces nœuds seront répartis de façon homogène. Le nombre de nœuds par défaut est 5.
- **Sélectionner les coordonnées des nœuds** : si vous activez cette option, vous devez sélectionner la plage contenant les coordonnées des nœuds

Onglet **Prédiction** :

Prédiction : activez cette option si vous souhaitez sélectionner des données à utiliser en mode prédiction. Si vous activez cette option, vous devez veiller à ce que les données de prédiction soient organisées comme les données d'estimation : mêmes variables, même ordre dans les sélections. En revanche vous ne devez pas sélectionner de libellés de variables : la première ligne des sélections décrites ci-dessous doit être une ligne de données.

Observations : sélectionner la ou les valeurs de X à prédire. La première ligne ne doit pas comprendre d'en-tête.

Libellés des observations : activez cette option si vous voulez utiliser des libellés d'observations disponibles sur une feuille Excel pour l'affichage des résultats. La première ligne ne doit pas comprendre d'en-tête. Si vous n'activez pas cette option, des libellés seront automatiquement créés (PredObs1, PredObs2, …).

Onglet **Données manquantes** :

Ne pas accepter les valeurs manquantes : activez cette option pour que XLSTAT empêche la poursuite des calculs si des valeurs manquantes sont détectées.

Supprimer les observations : activez cette option pour supprimer les observations comportant des données manquantes.

Onglet **Sorties**:

Statistiques descriptives : activez cette option pour afficher les statistiques descriptives pour les variables sélectionnées.

Corrélations : activez cette option pour afficher la matrice de corrélation entre les variables.

Prédictions et résidus : activez cette option pour afficher les prédictions et les résidus pour l'ensemble des observations.

Onglet **Graphiques** :

Courbe spline : activez cette option pour représenter la courbe spline

- **Prédictions et résidus** : activez cette option pour afficher les graphiques suivants :

(1) Variable explicative versus résidus normalisés : ce graphique n'est affiché que s'il n'y a qu'une seule variable explicative, et que cette variable est quantitative.

(2) Variable dépendante versus résidus normalisés.

(3) Prédictions pour la variable dépendante versus variable dépendante.

(4) Graphique en bâtons des résidus normalisés.

Résultats

Statistiques descriptives : dans ce tableau sont affichées les statistiques descriptives correspondantes aux 2 variables sélectionnées.

Coefficients de la spline cubique : les coefficients de la spline cubique pour chaque intervalle sont présentés dans un tableau.

Exemple

Un exemple d'application des splines cubiques est disponible sur le Centre d'aide XLSTAT :

http://www.xlstat.com/demo-splinesf.htm

Bibliographie

Guillod T. (2008). Interpolations, courbes de Bézier et B-splines. *Bulletin de la société des Enseignants Neuchatelois de Sciences*, **34**.

Régression non paramétrique

Cet outil permet de réaliser des régressions non paramétriques de deux types : *la Kernel regression* (régression par noyau) et la régression LOWESS.

Dans cette section :

Description

La régression non paramétrique peut être utilisée lorsque les hypothèses des méthodes de régression plus classiques ne sont pas vérifiées, ou lorsque la structure du modèle n'a pas fondamentalement d'intérêt et lorsque seule la qualité prédictive du modèle est importante.

Kernel regression:

La *Kernel regression* est un outil de modélisation faisant partie de la famille des méthodes de lissage. Contrairement à la régression linéaire qui est utilisée dans un but explicatif et prédictif (comprendre un phénomène pour pouvoir le prévoir ensuite), la *Kernel regression* est classée parmi les méthodes de régression non paramétrique essentiellement utilisées dans un but prédictif. La structure du modèle est en effet variable et complexe, ce dernier fonctionnant comme un filtre ou une boîte noire. De nombreuses variantes de la *Kernel regression* existent.

Comme pour toute méthode de modélisation, un échantillon d'apprentissage de taille n_{app} est utilisé pour estimer les paramètres du modèle. Un échantillon de validation de taille n_{valid} peut ensuite être utilisé pour évaluer la qualité du modèle. Enfin, le modèle peut être appliqué sur un échantillon de prédiction de taille n_{pred}, pour lequel les valeurs de la variable dépendante Y sont inconnues.

La première caractéristique de la *Kernel regression* est l'utilisation d'une fonction **noyau** pour pondérer les observations de l'échantillon d'apprentissage, en fonction de leur « distance » à l'observation prédite. Plus les valeurs des variables explicatives d'une observation de l'échantillon d'apprentissage sont proches des valeurs observées pour l'observation en cours de prédiction, plus le plus poids de l'observation de l'échantillon d'apprentissage sera important. Différents noyaux sont proposés dans la littérature scientifique. XLSTAT propose les noyaux suivants : Uniforme, Triangle, Epanechnikov, Quartic, Triweight, Tricube, Gaussien, et Cosinus.

La seconde caractéristique de la *Kernel regression* est la **bande passante** associée à chaque variable. Elle intervient dans le calcul et du noyau et du poids des observations, et permet de différencier ou d'homogénéiser le poids relatif des variables, tout en agissant sur l'impact d'une observation de l'échantillon d'apprentissage en fonction de sa distance à l'observation prédite. Le terme de bande passante fait allusion aux méthodes de filtrage. Pour une variable et une fonction noyau donnée, plus elle est faible, plus un nombre restreint d'observation influeront sur la prédiction.

Exemple : soit Y une variable dépendante, et k variables explicatives $(X_1, X_2, \ldots, X_k)$. Pour le calcul de la prédiction $y_i, (1 \leq i \leq n_{valid})$, étant donnée l'observation $j, (1 \leq j \leq n_{app})$, le poids déterminé par un noyau Gaussien avec une bande passante h_l pour chaque variable $X_l, (l = 1 \ldots k)$ est donné par :

$$w_{ij} = \frac{1}{\left(\sqrt{2\pi}\right)^k \prod_{l=1}^{k} h_l} \exp\left(-\sum_{l=1}^{k}\left(\frac{x_{jl} - x_{il}}{h_l}\right)^2\right)$$

La troisième caractéristique est le degré du modèle polynomial utilisé pour ajuster le modèle aux observations de l'échantillon d'apprentissage. Dans le cas du polynôme de degré 0 (polynôme constant), la formule de Nadaraya-Watson est utilisée pour calculer la prédiction i :

$$y_i = \frac{\sum_{j=1}^{n_{app}} w_{ij} y_j}{\sum_{j=1}^{n_{app}} w_{ij}}$$

Dans le cas du polynôme constant, les variables explicatives ne sont donc prises en compte que pour le calcul des poids des observations de l'échantillon d'apprentissage. Dans le cas des polynômes de degré 1 et 2 (la pratique montre que des ordres supérieurs ne sont pas nécessaires, et XLSTAT se limite aux degrés 0,1,2), les variables sont en revanche impliquées dans le calcul d'un modèle polynomial. Une fois le modèle calé, le modèle est appliqué aux observations des échantillons de validation et éventuellement de prédiction, afin d'estimer la valeur de la variable dépendante.

Une fois les paramètres du modèle estimés, on calcule la valeur de la prédiction en utilisant les formules suivantes :

1. Degré 1 : $y_i = a_0 + \sum_{l=1}^{k} a_l x_{il}^l$
2. Degré 2 : $y_i = a_0 + \sum_{l=1}^{k} a_l x_{il}^l + \sum_{l=1}^{k}\sum_{m=1}^{k} b_{lm} x_{il} x_{im}$

Remarques :

- Pour l'estimation des paramètres du polynôme, on pondère préalablement les observations de l'échantillon d'apprentissage en utilisant la formule de Nadaraya-Watson.

- Dans le cas d'un modèle d'ordre 1 ou 2, pour chaque observation des échantillons de validation et de prédiction, le modèle polynomial est estimé. La *Kernel regression* est donc une méthode potentiellement intensive.

Afin de limiter le nombre d'observations de l'échantillon d'apprentissage pris en compte pour l'estimation des paramètres du polynôme, plusieurs stratégies sont proposées :

- fenêtre glissante : pour l'estimation de la valeur y_i, on prend en compte un nombre fixé d'observations précédemment observées. Dans cette situation, l'échantillon d'apprentissage évolue donc en permanence.
- k plus proches voisins (*k nearest neighbors*) : cette méthode, éventuellement complémentaire de la précédente permet de limiter la taille de l'échantillon d'apprentissage à une valeur k donnée.

Détails concernant les fonctions noyau :

Pour calculer le poids w_ij de l'observation j pour le calcul de la prévision y_i, on définit :

$$W_{ij} = \prod_{l=1}^{k} \frac{K(u_{ijl})}{h_l} \text{ avec } u_{ijl} = \frac{x_{il} - x_{jl}}{h_l}$$

où K est une fonction noyau. Les différentes fonctions noyau proposées par XLSTAT sont :

- Uniforme : la fonction noyau est définie par :

$$K(u) = \frac{1}{2}\mathbb{I}_{|u|\leq 1}$$

- Triangle : la fonction noyau est définie par :

$$K(u) = (1 - |u|)\,\mathbb{I}_{|u|\leq 1}$$

- Epanechnikov : la fonction noyau est définie par :

$$K(u) = \frac{3}{4}\left(1 - u^2\right)\mathbb{I}_{|u|\leq 1}$$

- Quartic : la fonction noyau est définie par :

$$K(u) = \frac{15}{16}(1 - u^2)^2\mathbb{I}_{|u|\leq 1}$$

- Triweight : la fonction noyau est définie par :

$$K(u) = \frac{35}{32}(1 - u^2)^3\mathbb{I}_{|u|\leq 1}$$

- Tricube : la fonction noyau est définie par :

$$K(u) = \left(1 - |u|^3\right)^3\,\mathbb{I}_{|u|\leq 1}$$

- Gaussien : la fonction noyau est définie par :

$$K(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}}$$

- Cosinus : la fonction noyau est définie par :

$$K(u) = \frac{\pi}{4} cos\left(\frac{\pi}{2}u\right) \mathbb{I}_{|u|\leq 1}$$

Détails sur la régression LOWESS :

La régression LOWESS (Locally weighted regression and smoothing scatter plots) a été introduite par Cleveland (1979) dans le but de créer des courbes lissées passant au travers de nuages de points. De nouvelles versions ont depuis été mises au point afin d'augmenter la robustesse des modèles. La régression LOWESS est très proche de la *Kernel regression* car elle fait aussi appel à de la régression polynomiale avec des observations pondérées par une fonction noyau.

L'algorithme LOWESS peut être décrit comme suit : pour chaque individu i :

1 - Dans un premier temps on calcule les distances euclidiennes $d(i,j)$ entre l'individu i et l'individu j. Puis on sélectionne la fraction f des N individus les plus proches de i. Pour les points sélectionnés, on calcule leur poids en utilisant le noyau Tricube et la distance suivante :

$$D(i,j)\frac{d(i,j)}{Max_j\left(d(i,j)\right)}$$

$$Poids(j) = Tricube\left(D(i,j)\right)$$

2 - La régression est alors ajustée et une prévision est calculée pour l'individu i.

Pour la version robuste de la régression LOWESS, les étapes suivantes sont nécessaires :

3 - On recalcule les poids en utilisant la distance suivante :

$$D^{'}(i,j) = \frac{|r(j)|}{6Mediane_j\left(|r(j)|\right)}$$

où $r(j)$ est le résidu pour l'individu j à l'issue de l'étape précédente.

et en utilisant le noyau Quartic :

$$Poids(j) = Quartic\left(D^{'}(i,j)\right)$$

4 - La régression est alors ajustée de nouveau.

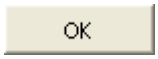
5 - on recommence les étapes 3 et 4. On obtient alors la prévision finale pour l'individu i.

Remarques :

- Hormis les données, les seuls paramètres d'entrées pour la méthode sont la fraction f d'individus les plus proches (exprimée en % dans XLSTAT) et l'ordre du polynôme.
- La *Robust LOWESS regression* est environ trois fois plus coûteuse en temps de calcul que la régression LOWESS.

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.

Onglet **Général**:

Y / Variables dépendantes :

Quantitatives : sélectionnez la ou les variables réponse que vous souhaitez modéliser. Si plusieurs variables sont sélectionnées, XLSTAT fera les calculs pour chacune des variables indépendamment. Si des en-têtes de colonnes ont été sélectionnés, veuillez vérifier que l'option « Libellés des variables » est activée.

X / Variables explicatives :

Quantitatives : activez cette option si vous voulez inclure une ou plusieurs variables explicatives quantitatives dans le modèle. Sélectionnez alors la ou les variables correspondantes sur la feuille Excel. Les données sélectionnées doivent être de type

numérique. Si le libellé des variables a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

Qualitatives : activez cette option si vous voulez inclure une ou plusieurs variables explicatives qualitatives dans le modèle. Sélectionnez alors la ou les variables correspondantes sur la feuille Excel. Les données sélectionnées peuvent être de tout type, mais les données numériques sont automatiquement considérées comme nominales. Si le libellé des variables a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des variables : activez cette option si la première ligne des données sélectionnées (variables dépendantes et explicatives, libellés des observations, poids) contient un libellé.

Libellés des observations : activez cette option si vous voulez utiliser des libellés d'observations pour l'affichage des résultats. Si l'option « Libellés des variables » est activée, la première cellule de la sélection doit comprendre un en-tête. Si vous n'activez pas cette option, des libellés seront automatiquement créés (Obs1, Obs2, …).

Méthode : choisissez la méthode de régression non paramétrique à utiliser (voir description).

Degré du polynôme : entrez le degré du polynôme dans le cas où la méthode choisie est la régression LOWESS ou un polynôme.

Onglet **Options**:

Echantillon d'apprentissage :

- **Fenêtre glissante** : choisissez cette option pour que la taille de l'échantillon d'apprentissage soit constante. Vous devez alors fixer la taille t de la fenêtre. Ainsi pour estimer la valeur $Y(i+1)$ les observations $i-t-1$ à i seront utilisées. La première observation pour laquelle une estimation sera calculée sera l'observation $t+1$.
- **Fenêtre croissante** : choisissez cette option pour que la taille de l'échantillon d'apprentissage soit croissante. Vous devez alors fixer la taille t de la fenêtre au départ. Ainsi pour estimer la valeur $Y(i+1)$ les observations 1 à i seront utilisées. La première observation pour laquelle une estimation sera calculée sera l'observation $t+1$.
- **Tout** : les échantillons d'apprentissage et de validation sont identiques. Si cette option n'a pas d'intérêt en matière de prévision, elle permet en revanche d'évaluer la méthode en

situation d'information parfaite.

K plus proches voisins : activez cette option pour définir la taille maximale de l'échantillon d'apprentissage. Deux options sont proposées :

- **Lignes** : les k points retenus seront les k points les plus proches du point à prédire, la proximité tenant compte de la bande passante.
- **%** : les points retenus seront les plus proches du point à prédire, et représenteront $x\%$ de l'échantillon d'apprentissage disponible, où x est la valeur à saisir.

Tolérance : entrez la valeur de la tolérance seuil en deçà de laquelle une variable est automatiquement ignorée.

Interactions / Niveau : activez cette option pour inclure des interactions dans le modèle puis entrez le niveau maximum d'interaction (valeur comprise entre 1 et 4).

Noyau : choisissez le type de fonction noyau à utiliser. Les options possibles sont : Uniforme, Triangle, Epanechnikov, Quartic, Triweight, Tricube, Gaussien, Cosinus. Une description de ces fonctions est disponible dans la partie description.

Bande passante : XLSTAT vous permet de choisir une méthode de calcul automatique de la bande passante ou de fixer les valeurs. Les différentes options possibles sont :

- **Constante** : la bande passante est constante et égale à la valeur fixée. Entrez alors la valeur de la bande passante.
- **Fixée** : la bande passante est définie dans une plage verticale de cellules Excel, que vous devez alors sélectionner. Le nombre de cellules doit être égal au nombre de variables explicatives, et les bandes passantes doivent être entrées dans le même ordre que les variables.
- **Amplitude** : la valeur de la bande passante h_l est déterminée pour la variable explicative X_l par la formule suivante :

$$h_l = Max\left(x_{il}\right)_{i=1,\ldots,n_{app}} - Min\left(x_{il}\right)_{i=1,\ldots,n_{app}}$$

- **Ecart-type** : la valeur de la bande passante h_l est égale, pour chaque variable explicative, à l'écart-type de la variable observée sur l'échantillon d'apprentissage.

Onglet **Validation** :

Validation : activez cette option si vous souhaitez utiliser une partie des données sélectionnées pour valider le modèle.

Jeu de validation : choisissez l'une des options pour définir le mode de sélection des observations utilisées pour la validation :

- **Aléatoire** : les observations sont sélectionnées de manière aléatoire. Le « Nombre d'observations » doit alors être saisi.

- **N dernières lignes** : les N dernières observations sont sélectionnées pour la validation. Le « Nombre d'observations » N doit alors être saisi.
- **N premières lignes** : les N premières observations sont sélectionnées pour la validation. Le « Nombre d'observations » N doit alors être saisi.
- **Variable de groupe** : si vous choisissez cette option, vous devez ensuite sélectionner une variable indicatrice composée de 0 pour les observations à utiliser pour le calcul du modèle, et de 1 pour les observations à utiliser pour la validation du modèle.

Onglet **Prédiction** :

Prédiction : activez cette option si vous souhaitez sélectionner des données à utiliser en mode prédiction. Si vous activez cette option, vous devez veiller à ce que les données de prédiction soient organisées comme les données d'estimation : mêmes variables, même ordre dans les sélections. En revanche vous ne devez pas sélectionner de libellés de variables : la première ligne des sélections décrites ci-dessous doit être une ligne de données.

Quantitatives : activez cette option pour sélectionner la ou les variables quantitatives explicatives. La première ligne ne doit pas comprendre d'en- tête.

Qualitatives : activez cette option pour sélectionner la ou les variables qualitatives explicatives. La première ligne ne doit pas comprendre d'en-tête.

Libellés des observations : activez cette option si vous voulez utiliser des libellés d'observations disponibles sur une feuille Excel pour l'affichage des résultats. La première ligne ne doit pas comprendre d'en-tête. Si vous n'activez pas cette option, des libellés seront automatiquement créés (PredObs1, PredObs2, …).

Libellés des variables : activez cette option si la première ligne des données sélectionnées (variables explicatives, libellés des observations) contient un libellé.

Onglet **Données manquantes** :

Supprimer les observations : activez cette option pour supprimer les observations comportant des données manquantes.

Estimer les données manquantes : activez cette option pour estimer les données manquantes avant le début des calculs.

- **Moyenne ou mode** : activez cette option pour estimer les données manquantes en utilisant la moyenne (variables quantitatives) ou le mode (variables qualitatives) pour les variables correspondantes.
- **Plus proche voisin** : activez cette option pour estimer les données manquantes d'une observation en recherchant le plus proche voisin de l'observation.

Onglet **Sorties**:

Statistiques descriptives : activez cette option pour afficher les statistiques descriptives pour les variables sélectionnées.

Corrélations : activez cette option pour afficher la matrice de corrélation des variables explicatives.

Coefficients d'ajustement : activez cette option pour afficher le tableau des statistiques d'ajustement du modèle.

Prédictions et résidus : activez cette option pour afficher les prédictions et les résidus pour l'ensemble des observations.

Onglet **Graphiques** :

Données et prédictions : activez cette option pour afficher le graphique des données observées et des prédictions:

- **En fonction de X1** : activez cette option pour afficher les valeurs observées et prédites en fonction des valeurs de la variable X_1.
- **En fonction du temps** : activez cette option pour sélectionner des données donnant la date correspondant à chacune des observations, afin d'afficher les résultats en fonction du temps.

Résidus : activez cette option pour afficher le diagramme en bâtons des résidus.

Résultats

Statistiques descriptives : le tableau de statistiques descriptives présente pour toutes les variables sélectionnées des statistiques simples. Pour les variables quantitatives sont affichés le nombre d'observations, le nombre de données manquantes, le nombre de données non manquantes, la moyenne, et l'écart-type (non biaisé). Pour les variables qualitatives, sont affichés les modalités, leurs effectifs et les pourcentages respectifs.

Matrice de corrélation : dans ce tableau sont affichées les corrélations entre les variables sélectionnées.

Coefficients d'ajustement : dans ce tableau sont affichées les statistiques suivantes :

- le coefficient de détermination R^2 ;
- la somme des carrés des erreurs (ou résidus) du modèle (SCE) ;
- la moyenne des carrés des erreurs (ou résidus) du modèle (MCE) ;
- la racine de la moyenne des carrés des erreurs (ou résidus) du modèle (RMCE) ;

Prédictions et résidus : ce tableau donne pour chaque observation les données de départ, la valeur prédite par le modèle et les résidus.

Graphiques :

Si une seule variable quantitative explicative à été sélectionnée, ou si une variable temporelle a été sélectionnée (option « en fonction du temps » de l'onglet « Graphiques » de la boîte de dialogue), le premier graphique représente les données et la courbe correspondant aux prédictions du modèle. Si l'option « en fonction de X1 » a été sélectionnée, le premier graphique correspond aux données observées et aux prédictions en fonction de la première variable explicative sélectionnée. Le second graphique affiché est le diagramme en bâtons des résidus.

Exemple

Un exemple de *Kernel regression* est disponible sur le Centre d'aide XLSTAT à l'adresse

http://www.xlstat.com/demo-kernelf.htm

Bibliographie

Cleveland W.S. (1979). Robust locally weighted regression and smoothing scatterplots. *J. Amer. Statist. Assoc.*, **74**, 829-836.

Cleveland W.S. (1994). The Elements of Graphing Data. Hobart Press, Summit, New Jersey.

Härdle W. (1992). Applied Nonparametric Regression. Cambridge University Press, Cambridge.

Nadaraya E.A. (1964). On estimating regression.*Theory Probab. Appl.*, **9**, 141-142.

Wand M.P. and Jones M.C. (1995). Kernel Smoothing. Chapman and Hall, New York.

Watson G.S. (1964). Smooth regression analysis.*Sankhyâ Ser.A*, **26**, 101-116.

Régression non linéaire

Utilisez cet outil pour ajuster des données à n'importe quelle fonction linéaire ou non linéaire. La méthode utilisée est celle des moindres carrés. Il est possible d'utiliser soit des fonctions préprogrammées, soit des fonctions ajoutées par l'utilisateur.

Dans cette section :

Description

La régression non linéaire permet de modéliser des phénomènes complexes n'entrant pas dans le cadre du modèle linéaire. XLSTAT propose des fonctions préprogrammées parmi lesquelles l'utilisateur pourra éventuellement trouver le modèle décrivant le phénomène à modéliser.

Lorsque le modèle recherché n'est pas disponible, l'utilisateur a la possibilité de définir un nouveau modèle et de l'ajouter à sa librairie personnelle. Pour améliorer la vitesse et la fiabilité des calculs, il est recommandé d'ajouter les dérivées de la fonction par rapport à chacun des paramètres du modèle.

Afin de calculer ces paramètres, l'algorithme de Levenberg-Marquardt est utilisé.

XLSTAT permet de modéliser plusieurs fonctions à la fois. Vous pouvez ensuite choisir d'afficher uniquement les résultats du meilleur modèle (en se basant sur l'AIC), ou ceux de l'ensemble des modèles.

Algorithme de Levenberg-Marquardt

L'algorithme de Levenberg-Marquardt est une technique itérative qui localise le minimum d'une fonction multivariée. Cette technique est recommandée pour les problèmes de régression non-linéaire des moindres carrés.

Considérons le modèle non linéaire $Y = f(X, \Theta)$, où θ est le vecteur de paramètres de taille px_1, X est le vecteur des variables explicatives, et f est une fonction de X et theta. Le but est de trouver l'estimation des moindres carrés Θ' de Θ telle que Θ' minimise la fonction f.

On cherche donc à minimiser la fonction suivante :

$$g(\Theta) = \sum_{i=1}^{m} \left(Y_i - f(\Theta, X_i)\right)^2) \cdot$$

La procédure est itérative, on part d'un paramètre initial, si possible proche de la solution finale, et on applique la routine suivante :

$$\Theta_{j+1} = \Theta_j - (J'J + \lambda D)^{-1} J'(Y - f(\Theta, X_i)),$$

où J est la matrice Jacobienne, et D est une matrice diagonale pour ajuster le paramètre d'amortissement λ.

Lorsque les dérivées des fonctions ne sont pas spécifiées, leurs estimations sont faites via une approximation par différences finies.

Ajustement global et paramètres partagés

XLSTAT offre la possibilité d'ajuster plusieurs variables en même temps. Pour cela, deux options au choix sont disponibles :

- La première dans le cas où il y a une colonne pour chaque variable Y à ajuster.
- La seconde dans le cas où il y a une colonne contenant l'ensemble des variables Y à ajuster ainsi qu'une colonne d'indices de groupe permettant d'identifier chaque Y. Avec cette option, il y a la possibilité de choisir un ensemble de paramètres partagés s'appliquant à un ensemble de courbes.

Ajouter une fonction à la librairie des fonctions définies par l'utilisateur

Syntaxe :

Les paramètres doivent être représentés sous la forme pr_1, pr_2, …

Les variables explicatives doivent être représentées sous la forme X_1, X_2, …

Les fonctions Excel peuvent être utilisées : Exp(), Sin(), Pi(), Max()…

Exemple de fonction : pr_1 * Exp(pr_2 + pr_3 * X_1 + pr_4 * X_2)

Fichier contenant les définitions de fonction :

La librairie des fonctions utilisateur est enregistrée dans le fichier Models.txt, dans le répertoire utilisateur, tel qu'il est défini lors de l'installation ou au travers de la boîte des options de XLSTAT. Cette librairie est construite de la façon suivante :

Ligne 1 : nombre de fonctions définies par l'utilisateur.

Ligne 2 : N1 = nombre de paramètres intervenant dans la fonction 1.

Ligne 3 : définition de la fonction 1.

Lignes 4 à (3 + N1) : définition des dérivées de la fonction 1.

Ligne 4+N1 : N2= nombre de paramètres intervenant dans la fonction 2.

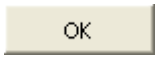
Ligne 5+N1 : définition de la fonction 2…

Lorsque les dérivées sont inconnues, « Unknown » remplace chaque dérivée de la fonction.

Vous pouvez modifier manuellement les éléments de ce fichier mais veillez à ne pas introduire d'erreur.

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.


: cliquez sur ce bouton pour changer la façon dont vous chargez les données dans XLSTAT. Si le bouton possède une icône de souris, XLSTAT vous permet de faire une sélection à la souris des données. Si le bouton possède une icône de liste, XLSTAT vous permet de sélectionner les données à partir d'une liste. Si le bouton possède une icône orange, des boutons complémentaires avec un point d'interrogation  sont affichés pour vous permettre d'importer les données à partir de fichiers.

Onglet **Général** :

Y / Variables dépendantes :

Quantitatives : sélectionnez la ou les variables réponse que vous souhaitez modéliser. Si plusieurs variables sont sélectionnées, XLSTAT fera les calculs pour chacune des variables indépendamment. Si des en-têtes de colonnes ont été sélectionnés, veuillez vérifier que l'option « Libellés des variables » est activée.

X / Variables explicatives :

Quantitatives : activez cette option si vous voulez inclure une ou plusieurs variables explicatives quantitatives dans le modèle. Sélectionnez alors la ou les variables

correspondantes sur la feuille Excel. Les données sélectionnées doivent être de type numérique. Si le libellé des variables a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

Groupes : activez cette option si vous souhaitez inclure une variable de groupes. Celle-ci vous permettra d'ajuster plusieurs variables en même temps. Sélectionnez alors la variable correspondante sur la feuille Excel. Si le libellé des variables a été sélectionné, vérifiez également que l'option « Libellés des variables » est également activée.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des variables : activez cette option si la première ligne des données sélectionnées (variables dépendantes et explicatives, libellés des observations, poids) contient un libellé.

Libellés des observations : activez cette option si vous voulez utiliser des libellés d'observations pour l'affichage des résultats. Si l'option « Libellés des variables » est activée, la première cellule de la sélection doit comprendre un en-tête. Si vous n'activez pas cette option, des libellés seront automatiquement créés (Obs1, Obs2, …).

Poids des observations : activez cette option si vous voulez pondérer les observations. Si vous n'activez pas cette option, les poids seront tous considérés comme valant 1. XLSTAT prend en compte ces poids pour les calculs des degrés de libertés. Les poids doivent être impérativement supérieurs ou égaux à 0. Si un en-tête de colonne a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

Onglet **Fonctions** :

Fonctions préprogrammées : activez cette option pour ajuster aux données l'une des fonctions disponibles dans la liste des fonctions préprogrammées. Sélectionnez alors une fonction dans la liste.

Fonctions définies par l'utilisateur : activez cette option pour ajuster aux données l'une des fonctions disponibles dans la liste des fonctions définies par l'utilisateur, ou pour ajouter une nouvelle fonction.

Choisir un modèle : activez cette option pour ajuster aux données une fonction et afficher ses résultats.

Choisir un modèle parmi plusieurs : activez cette option pour ajuster aux données plusieurs fonctions et afficher les résultats du meilleur modèle, en se basant sur l'AIC.

Choisir plusieurs modèles : activez cette option pour ajuster aux données plusieurs fonctions et afficher leurs résultats.

Editer : cliquez sur ce bouton pour faire apparaître dans la zone « Fonction : Y = » la fonction préprogrammée active. Vous pourrez alors copier la fonction pour ensuite la modifier pour créer une nouvelle fonction ou les dérivées d'une nouvelle fonction.

Supprimer : cliquez sur ce bouton pour supprimer la fonction active de la liste des fonctions définies par l'utilisateur.

Ajouter : cliquez sur ce bouton pour ajouter une fonction à la liste des fonctions définies par l'utilisateur. Vous devez alors entrer la fonction dans le champ « **Fonction : Y =** », puis, si vous le souhaitez, sachant que cela permet d'améliorer la vitesse des calculs et d'obtenir les écarts types des paramètres, vous pouvez sélectionner les dérivées de la fonction par rapport à chacun des paramètres. Pour cela activez l'option « **Dérivées** », puis sélectionnez sur une feuille Excel les dérivées.

Dérivées : sélectionnez sur une feuille Excel les dérivées, sachant que cela permet d'améliorer la vitesse des calculs.

Détails : Cliquer sur ce bouton pour obtenir des informations supplémentaires sur la fonction préprogrammée sélectionnée.

Concentration de l'inhibiteur : dans le cas ou vous sélectionnez au moins une des 4 fonctions suivantes : "Inhibition competitive", "Inhibition non compétitive", "Inhibition incompétitive", ou "Inhibition de modèle mixte", vous devez sélectionner la concentration de l'inhibiteur. Il vous faut alors une concentration différente par groupe.

Remarque : la section description contient des informations relatives à la définition des fonctions utilisateur.

Onglet **Options**:

Valeurs de départ : activez cette option pour donner un point de départ à XLSTAT. Sélectionnez alors les cellules correspondant aux valeurs initiales des paramètres. Le nombre de lignes sélectionnées doit correspondre au nombre de paramètres.

Bornes des paramètres : activez cette option pour indiquer à XLSTAT une région possible pour l'ensemble des paramètres du modèle choisi. Vous devez alors sélectionner une plage de deux colonnes, celle de gauche correspondant aux bornes inférieures, et celle de droite aux bornes supérieures. Le nombre de lignes sélectionnées doit correspondre au nombre de paramètres.

Libellés des paramètres : activez cette option si vous voulez préciser les noms des paramètres. Au lieu d'afficher les noms génériques pr1, pr2, etc., pour les paramètres, XLSTAT affichera les résultats en utilisant les libellés sélectionnés. Le nombre de lignes sélectionnées doit correspondre au nombre de paramètres.

Paramètres partagés : activez cette option si vous souhaitez ajouter des paramètres partagés dans le modèle. Cette option n'est uniquement disponible que dans le cas où vous avez ajouté une variable de groupes. Les paramètres partagés auront alors la même valeur de paramètre pour l'ensemble des groupes ajustés.

Conditions d'arrêt :

- **Itérations** : entrez le nombre maximal d'itérations pour l'algorithme d'ajustement. Les calculs sont interrompus dès que le nombre maximal d'itérations est dépassé. Valeur par défaut : 50.

- **Convergence** : entrez la valeur seuil d'évolution maximale de la somme des carrés des erreurs (SCE) d'une itération à l'autre, qui une fois atteinte permet de considérer que l'algorithme a convergé. Valeur par défaut : 0,0001.

Onglet **Validation** :

Validation : activez cette option si vous souhaitez utiliser une partie des données sélectionnées pour valider le modèle.

Jeu de validation : choisissez l'une des options pour définir le mode de sélection des observations utilisées pour la validation :

- **Aléatoire** : les observations sont sélectionnées de manière aléatoire. Le « Nombre d'observations » doit alors être saisi.
- **N dernières lignes** : les N dernières observations sont sélectionnées pour la validation. Le « Nombre d'observations » N doit alors être saisi.
- **N premières lignes** : les N premières observations sont sélectionnées pour la validation. Le « Nombre d'observations » N doit alors être saisi.
- **Variable de groupe** : si vous choisissez cette option, vous devez ensuite sélectionner une variable indicatrice composée de 0 pour les observations à utiliser pour le calcul du modèle, et de 1 pour les observations à utiliser pour la validation du modèle.

Onglet **Prédiction** :

Prédiction : activez cette option si vous souhaitez sélectionner des données à utiliser en mode prédiction. Si vous activez cette option, vous devez veiller à ce que les données de prédiction soient organisées comme les données d'estimation : mêmes variables, même ordre dans les sélections. En revanche vous ne devez pas sélectionner de libellés de variables : la première ligne des sélections décrites ci-dessous doit être une ligne de données.

Quantitatives : activez cette option pour sélectionner la ou les variables quantitatives explicatives. La première ligne ne doit pas comprendre d'en-tête.

Qualitatives : activez cette option pour sélectionner la ou les variables qualitatives explicatives. La première ligne ne doit pas comprendre d'en-tête.

Libellés des observations : activez cette option si vous voulez utiliser des libellés d'observations disponibles sur une feuille Excel pour l'affichage des résultats. La première ligne ne doit pas comprendre d'en-tête. Si vous n'activez pas cette option, des libellés seront automatiquement créés (PredObs1, PredObs2, …).

Libellés des variables : activez cette option si la première ligne des données sélectionnées (variables explicatives, libellés des observations) contient un libellé.

Onglet **Données manquantes** :

Supprimer les observations : activez cette option pour supprimer les observations comportant des données manquantes.

Estimer les données manquantes : activez cette option pour estimer les données manquantes avant le début des calculs.

- **Moyenne ou mode** : activez cette option pour estimer les données manquantes en utilisant la moyenne (variables quantitatives) ou le mode (variables qualitatives) pour les variables correspondantes.
- **Plus proche voisin** : activez cette option pour estimer les données manquantes d'une observation en recherchant le plus proche voisin de l'observation.

Onglet **Sorties** :

Statistiques descriptives : activez cette option pour afficher les statistiques descriptives pour les variables sélectionnées.

Corrélations : activez cette option pour afficher la matrice de corrélations des variables explicatives.

Coefficients d'ajustement : activez cette option pour afficher le tableau des statistiques d'ajustement du modèle.

Paramètres du modèle : activez cette option pour afficher la valeur des paramètres du modèle après ajustement.

Prédictions et résidus : activez cette option pour afficher les prédictions et les résidus pour l'ensemble des observations.

Graphiques :

- **Données et prédictions** : activez cette option pour afficher le graphique des données observées et la courbe de la fonction ajustée.
- **Résidus** : activez cette option pour afficher le diagramme en bâtons des résidus.

Résultats

Statistiques simples : le tableau des statistiques descriptives présente pour toutes les variables sélectionnées des statistiques simples : le nombre d'observations, le nombre de données manquantes, le nombre de données non manquantes, la moyenne, et l'écart-type (non biaisé).

Matrice de corrélation : dans ce tableau sont affichées les corrélations entre les variables sélectionnées.

Tableau de synthèse : ce tableau est affiché dans le cas où on a choisi d'ajuster plusieurs modèles. Dans ce tableau sont affichés les coefficients d'ajustement pour chacun des modèles ajustés. A la suite de ce tableau est affiché un graphique des AIC de chacun des modèles.

Coefficients d'ajustement : dans ce tableau sont affichées les statistiques suivantes :

- le nombre d'observations ;
- le nombre de degrés de liberté (DDL) ;

- le coefficient de détermination R^2 ;
- la somme des carrés des erreurs (ou résidus) du modèle (SCE) ;
- la moyenne des carrés des erreurs (ou résidus) du modèle (MCE) ;
- la racine de la moyenne des carrés des erreurs (ou résidus) du modèle (RMCE) ;
- **AIC** : le critère d'information d'Akaike (Akaike's Information Criterion) ;
- **AICC** : le critère d'information d'Akaike Corrigé (Akaike's Information Criterion Corrected) ;
- Le nombre d'itérations avant convergence.

Paramètres du modèle : ce tableau fournit pour chaque paramètre sa valeur après ajustement du modèle, son écart-type associé ainsi que l'intervalle de confiance à 95%.

Prédictions et résidus : ce tableau donne pour chaque observation les données de départ, la valeur prédite par le modèle et les résidus.

Graphiques :

Si une seule variable quantitative explicative à été sélectionnée le premier graphique représente les données et la courbe correspondant à la fonction choisie. Dans ce graphique, il vous est possible d'afficher les courbes d'intervalle de confiance à 95% ainsi que les courbes d'intervalle de prévision à 95%. L'intervalle de confiance vous permet d'évaluer la valeur ajustée pour les valeurs observées des variables, alors que l'intervalle de prévision donne une étendue des valeurs autour de laquelle une observation future de la variable dépendante peut être attendue.

Le second graphique affiché est le diagramme en bâtons des résidus.

Exemple

Des exemples de régression non linéaire sont disponibles sur le Centre d'aide XLSTAT à l'adresse :

http://www.xlstat.com/demo-nonlinf.htm http://www.xlstat.com/demo-nonlinf2.htm

Bibliographie

Ramsay J.O. and Silverman B.W. (1997). Functional Data Analysis. Springer-Verlag, New York.

Ramsay J.O. and Silverman B.W. (2002). Applied Functional Data Analysis. Springer-Verlag, New York.

Doubles moindres carrés (2SLS)

Utilisez cet outil pour analyser vos données par la méthode des doubles moindres carrés.

Dans cette section :

Description

La méthode des doubles moindres carrés (*2SLS ou Two-Stage Least Squares regression* en anglais) est très utilisée lorsque, dans une régression linéaire, au moins une des variables explicatives est endogène. Dans ce cas, la variable sera corrélée au terme d'erreur, ce qui est en contradiction avec les hypothèses de la régression linéaire. On peut notamment rencontrer ce type de situation lorsque l'une des variables explicatives a été mesurée avec erreur

Le principe de la méthode des doubles moindres carrés est d'utiliser des variables instrumentales non corrélées au terme d'erreur pour estimer les paramètres du modèle. Ces variables instrumentales sont des variables corrélées aux variables endogènes mais pas à leur terme d'erreur.

On note y une variable dépendante quantitative, X_1 la matrice des $p1$ variables explicatives endogènes (corrélées au terme d'erreur), X_2 la matrice des $p2$ variables explicatives exogènes (non corrélées au terme d'erreur) ($p = p1 + p2$) et Z la matrice des q variables instrumentales (corrélées à X_1 mais pas au terme d'erreur).

Les équations structurelles s'écrivent :

$$\begin{cases} y = X_1\beta_1 + X_2\beta_2 + \epsilon \\ X_1 = Z\gamma + \delta \end{cases}$$

où β_1 et β_2 représentent respectivement les paramètres associés à X_1 et X_2, γ le paramètre de la régression entre X_1 et Z. Les variables ϵ et δ correspondent aux termes d'erreurs.

A partir de la méthode d'estimation proposée dans (Theil, 1953a), (Theil, 1953b), l'estimateur du paramètre est donné par :

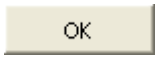
$$\hat{\beta} = (X'\Omega X)^{-1} X'\Omega X y$$

avec Ω la matrice de projection définie telle que $\Omega = Z(Z'Z)^{-1}Z'$.

XLSTAT permet de prendre en compte X_1, X_2 et$Z$. Pour cela, sélectionnez X_1 et X_2 en tant que variables explicatives et Z et X_2 en tant que variables instrumentales (X_2 est sélectionnée deux fois). Vous obtiendrez alors les coefficients β_1 et β_2.

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.

Onglet **Général**:

Y / Variables dépendantes :

Quantitatives : sélectionnez la ou les variable(s) réponse(s) que vous souhaitez modéliser. Si plusieurs variables sont sélectionnées, XLSTAT fera les calculs pour chacune des variables indépendamment. Si des en-têtes de colonnes ont été sélectionnés, veuillez vérifier que l'option « Libellés des variables » est activée.

X / Variables explicatives :

Quantitatives : sélectionnez la ou les variable(s) quantitative(s) explicative(s) sur la feuille Excel. Les données sélectionnées doivent être de type numérique. Si le libellé des variables a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée. Il faut sélectionner aussi bien les variables explicatives endogènes qu'exogènes.

Z / Variables instrumentales :

Quantitatives : sélectionnez la ou les variable(s) quantitative(s) instrumentale(s) sur la feuille Excel. Les données sélectionnées doivent être de type numérique. Si le libellé des variables a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée. Les variables explicatives exogènes doivent aussi figurer parmi les variables instrumentales.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des variables : activez cette option si la première ligne des données sélectionnées (variables dépendantes et explicatives, libellés des observations, poids) contient un libellé.

Libellés des observations : activez cette option si vous voulez utiliser des libellés d'observations pour l'affichage des résultats. Si l'option « Libellés des variables » est activée, la première cellule de la sélection doit comprendre un en-tête. Si vous n'activez pas cette option, des libellés seront automatiquement créés (Obs1, Obs2, …).

Poids des observations : activez cette option si vous voulez pondérer les observations. Si vous n'activez pas cette option, les poids seront tous considérés comme valant 1. Les poids doivent être impérativement supérieurs ou égaux à 0. Un poids de 2 est équivalent à répéter deux fois la même observation. Si un en-tête de colonne a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

Poids dans la régression : activez cette option si vous voulez effectuer une régression par les moindres carrés pondérés. Si vous n'activez pas cette option, les poids seront tous considérés comme valant 1. Les poids doivent être impérativement supérieurs ou égaux à 0. Si un en-tête de colonne a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

Onglet **Options**:

Constante nulle : activez cette option pour exclure la constante du modèle.

Tolérance : activez cette option pour permettre à l'algorithme de calcul de la régression OLS de ne pas prendre en compte les variables qui seraient soit constantes soit trop corrélées avec d'autres variables déjà utilisées dans le modèle (0.0001 par défaut).

Intervalle de confiance (%) : entrez l'étendue en pourcentage de l'intervalle de confiance à utiliser pour les différents tests, et pour le calcul des intervalles de confiance autour des paramètres et des prédictions. Valeur par défaut : 95.

Onglet **Validation** :

Validation : activez cette option si vous souhaitez utiliser une partie des données sélectionnées pour valider le modèle.

Jeu de validation : choisissez l'une des options pour définir le mode de sélection des observations utilisées pour la validation :

- **Aléatoire** : les observations sont sélectionnées de manière aléatoire. Le « Nombre d'observations » doit alors être saisi.
- **N dernières lignes** : les N dernières observations sont sélectionnées pour la validation. Le « Nombre d'observations » N doit alors être saisi.
- **N premières lignes** : les N premières observations sont sélectionnées pour la validation. Le « Nombre d'observations » N doit alors être saisi.
- **Variable de groupe** : si vous choisissez cette option, vous devez ensuite sélectionner une variable indicatrice composée de 0 pour les observations à utiliser pour le calcul du modèle, et de 1 pour les observations à utiliser pour la validation du modèle.

Onglet **Prédiction** :

Prédiction : activez cette option si vous souhaitez sélectionner des données à utiliser en mode prédiction. Si vous activez cette option, vous devez veiller à ce que les données de prédiction soient organisées comme les données d'estimation : mêmes variables, même ordre dans les sélections. En revanche vous ne devez pas sélectionner de libellés de variables : la première ligne des sélections décrites ci-dessous doit être une ligne de données.

X / Variables explicatives : sélectionner la ou les variables quantitatives explicatives. La première ligne ne doit pas comprendre d'en- tête. Il faut s sélectionner les variables explicatives exogènes et endogènes. Il n'est pas nécessaire de sélectionner les variables instrumentales.

Libellés des observations : activez cette option si vous voulez utiliser des libellés d'observations disponibles sur une feuille Excel pour l'affichage des résultats. La première ligne ne doit pas comprendre d'en-tête. Si vous n'activez pas cette option, des libellés seront automatiquement créés (PredObs1, PredObs2, …).

Onglet **Données manquantes** :

Supprimer les observations : activez cette option pour supprimer les observations comportant des données manquantes.

- **Vérifier pour chaque Y séparément** : choisissez cette option si vous voulez que lorsque, pour une observation donnée, il y a des données manquantes uniquement dans les Y, l'observation ne soit supprimée que si la donnée correspondant au Y en cours de modélisation est manquante.
- **Pour tous les Y** : choisissez cette option pour supprimer toutes les observations pour lesquelles des Y sont manquants.
- Remarque : les deux alternatives ci-dessus sont sans effet si il n'y a qu'un seul Y.

Estimer les données manquantes : activez cette option pour estimer les données manquantes avant le début des calculs.

- **Moyenne ou mode** : activez cette option pour estimer les données manquantes en utilisant la moyenne (variables quantitatives) ou le mode (variables qualitatives) pour les variables correspondantes.
- **Plus proche voisin** : activez cette option pour estimer les données manquantes d'une observation en recherchant le plus proche voisin de l'observation.

Onglet **Sorties**:

Statistiques descriptives : activez cette option pour afficher les statistiques descriptives pour les variables sélectionnées.

Corrélations : activez cette option pour afficher la matrice de corrélation pour les variables quantitatives (dépendantes et explicatives).

Analyse de la variance : activez cette option pour afficher le tableau d'analyse de la variance.

Coefficients normalisés : activez cette option pour afficher les paramètres normalisés du modèle (coefficients bêta).

Prédictions et résidus : activez cette option pour afficher les prédictions et les résidus pour l'ensemble des observations.

Onglet **Graphiques** :

Options communes :

Graphiques de régression : activez cette option pour afficher les graphiques de régression :

- **Coefficients normalisés** : activez cette option pour afficher sur un graphique les paramètres normalisés du modèle avec leur intervalle de confiance.
- **Prédictions et résidus** : activez cette option pour afficher les graphiques suivants :

(1) Droite de régression : ce graphique n'est affiché que s'il n'y a qu'une seule variable explicative, et que cette variable est quantitative.

(2) Variable explicative versus résidus normalisés : ce graphique n'est affiché que s'il n'y a qu'une seule variable explicative, et que cette variable est quantitative.

(3) Variable dépendante versus résidus normalisés.

(4) Prédictions pour la variable dépendante versus variable dépendante.

(5) Graphique en bâtons des résidus normalisés.

- **Intervalles de confiance** : activez cette option pour afficher les intervalles de confiance sur les graphiques (1) et (4).

Résultats

Statistiques descriptives : les tableaux de statistiques descriptives présentent pour toutes les variables sélectionnées des statistiques simples. Pour les variables dépendantes (en bleu), les variables explicatives et les variables instrumentales, sont affichés le nombre d'observations, le nombre de données manquantes, le nombre de données non manquantes, la moyenne, et l'écart-type (non biaisé).

Matrice de corrélation : ce tableau est affiché afin de vous permettre d'avoir un aperçu des corrélations entre les différentes variables sélectionnées.

Coefficients d'ajustement : dans ce tableau sont affichées les statistiques relatives à l'ajustement du modèle de régression :

- **Observations** : le nombre d'observations prises en compte dans les calculs. Dans les formules présentées ci-dessous n désigne le nombre d'observations.
- **Somme des poids** : la somme des poids des observations prises en compte dans les calculs. Dans les formules présentées ci-dessous W désigne la somme des poids.
- **DDL** : le nombre de degrés de liberté pour le modèle retenu (correspondant à la partie erreurs).
- **R²** : le coefficient de détermination du modèle. Ce coefficient, dont la valeur est comprise entre 0 et 1, n'est affiché que si la constante du modèle n'est pas fixée par l'utilisateur. Sa valeur est définie par

$$\mathrm{R}^2 = 1 - \frac{\sum_{i=1}^{n} w_i \left(y_i - \hat{y}_i\right)^2}{\sum_{i=1}^{n} w_i \left(y_i - \bar{y}\right)^2} \quad \text{avec } \bar{y} = \frac{1}{n}\sum_{i=1}^{n} w_i y_i$$

- Le R² s'interprète comme la proportion de la variabilité de la variable dépendante expliquée par le modèle. Plus le R² est proche de 1, meilleur est le modèle. L'inconvénient du R² est qu'il ne prend pas en compte le nombre de variables utilisées pour ajuster le modèle.
- **R²ajusté** : le coefficient de détermination ajusté du modèle. Le R² ajusté peut être négatif si le R² est voisin de zéro. Ce coefficient n'est affiché que si la constante du modèle n'est pas fixée par l'utilisateur. Sa valeur est définie par

$$\hat{R}^2 = 1 - (1 - R^2)\frac{W - 1}{W - p - 1}$$

Le R² ajusté est une correction du R² qui permet de prendre en compte le nombre de variables utilisées dans le modèle.

- **MCE** : la moyenne des carrés des erreurs (MCE) est définie par :

$$MCE = \frac{1}{W - p^*} \sum_{i=1}^{n} w_i (y_i - \hat{y}_i)^2$$

- **RMCE** : la racine de la moyenne des carrés des erreurs (RMCE) est la racine carrée de la MCE.

- **MAPE** : la *Mean Absolute Percentage Error* est calculée comme suit :

$$MAPE = \frac{100}{W} \sum_{i=1}^{n} w_i \left| \frac{y_i - \hat{y}_i}{y_i} \right|$$

- **DW** : le coefficient de Durbin-Watson est défini par

$$DW = \frac{\sum_{i=2}^{x} [(y_i - \hat{y}_i) - (y_{i-1} - \hat{y}_{i-1})]^2}{\sum_{i=1}^{x} w_i (y_i - \hat{y}_i)^2}$$

Ce coefficient correspond au coefficient d'autocorrélation d'ordre 1 et permet de vérifier que les résidus du modèle ne sont pas autocorrélés, sachant que l'indépendance des résidus est l'une des hypothèses de base de la régression linéaire. L'utilisateur pourra se référer à une table des coefficients de Durbin-Watson pour vérifier si l'hypothèse d'indépendance des résidus est acceptable.

- **Cp** : le coefficient Cp de Mallows est défini par

$$Cp = \frac{SCE}{\hat{\sigma}} + 2p^* - W$$

où SCE est la somme du carré des erreurs pour le modèle avec p variables explicatives, et où $\hat{\sigma}$ correspond à l'estimateur de la variance des résidus pour le modèle comprenant toutes les variables explicatives. Plus le coefficient Cp est proche de p* moins le modèle est biaisé.

- **AIC** : le critère d'information d'Akaike (Akaike's Information Criterion) est défini par

$$AIC = W \ln(\frac{SCE}{W}) + 2p^*$$

Ce critère proposé par Akaike (1973) dérive de la théorie de l'information, et s'appuie sur la mesure de Kullback et Leibler (1951). C'est un critère de sélection de modèles qui pénalise les modèles pour lesquels l'ajout de nouvelles variables explicatives n'apporte pas suffisamment d'information au modèle, l'information étant mesurée au travers de la SCE. On cherche à minimiser le critère AIC.

- **SBC** : le critère bayésien de Schwarz (Schwarz's Bayesian Criterion) est défini par

$$SBC = W \ln(\frac{SCE}{W}) + \ln(W)p^*$$

Ce critère proposé par Schwarz (1978) est proche du critère AIC, et comme ce dernier on cherche à le minimiser.

- **PC** : le critère de prédiction d'Amemiya (Amemiya's Prediction Criterion) est défini par

$$PC = \frac{(1 - R^2)(W + p^*)}{W - p^*}$$

Ce critère proposé par Amemiya (1980) permet comme le R² ajusté de tenir compte de la parcimonie du modèle.

- **Press RMCE** : la statistique de Press n'est affichée que si l'option correspondante a été activée dans la boîte de dialogue. Elle est définie par

$$Press = \sum_{i=1}^{n} w_i (y_i - \hat{y}_{i(-i)})^2$$

où $\hat{y}_{i(-i)}$ est la prédiction pour l'observation i lorsque cette dernière n'est pas utilisée pour l'estimation des paramètres. On obtient alors

$$\text{Press RMCE } = \sqrt{\frac{\text{Press}}{W - p^*}}$$

Le Press RMCE peut alors être comparé au RMCE. Une différence importante entre les deux indique que le modèle est sensible à la présence ou absence de certaines observations dans le modèle.

Le **tableau d'analyse de la variance** permet d'évaluer le pouvoir explicatif des variables explicatives. Dans le cas où la constante du modèle n'est pas fixée à une valeur donnée, le pouvoir explicatif est évalué en comparant l'ajustement (au sens des moindres carrés) du modèle final avec l'ajustement du modèle rudimentaire composé d'une constante égale à la moyenne de la variable dépendante. Dans le cas où la constante du modèle est fixée, la comparaison est faite par rapport au modèle pour lequel la variable dépendante serait égale à la constante fixée.

L'**équation du modèle** est ensuite affichée pour faciliter la lecture ou la réutilisation du modèle.

Le tableau des **coefficients normalisés** (aussi appelés coefficients bêta) permet de comparer le poids relatif des variables. Plus la valeur absolue d'un coefficient est élevée, plus le poids de la variable correspondante est important. Lorsque l'intervalle de confiance autour des coefficients normalisés comprend la valeur 0 (cela est facilement visible sur le graphique des coefficients normalisés), le poids d'une variable dans le modèle n'est pas significatif.

Dans le tableau des **prédictions et résidus** sont donnés pour chaque observation, son poids, la valeur de la variable explicative qualitative s'il n'y en a qu'une, la valeur observée de la variable dépendante, la prédiction du modèle, les résidus, les intervalles de confiance et la prédiction ajustée. Deux types d'intervalles de confiance sont affichés : un intervalle de confiance autour de la moyenne (correspondant au cas où l'on ferait la prédiction pour un nombre infini d'observations avec un ensemble de valeurs données des variables explicatives) et un intervalle autour de la prédiction ponctuelle (correspondant au cas d'une prédiction isolée pour des valeurs données des variables explicatives). Le second intervalle est toujours plus grand que le premier, les aléas étant plus importants. Si des données de validation ont été sélectionnées, elles sont affichées en fin de tableau.

Les **graphiques** qui suivent permettent de visualiser les résultats mentionnés ci-dessus. S'il n'y a qu'une seule variable explicative dans le modèle, le premier graphique affiché permet de visualiser les valeurs observées, la droite de régression et les deux types d'intervalles de confiance autour des prévisions. Le second graphique permet quant à lui de visualiser les résidus normalisés en fonction de la variable explicative. En principe, les résidus doivent être distribués de manière aléatoire autour de l'axe des abscisses. L'observation d'une tendance ou d'une forme révèlerait un problème au niveau du modèle.

Les **trois graphiques** affichés ensuite permettent de visualiser respectivement l'évolution des résidus normalisés en fonction de la variable dépendante, la distance entre les prédictions et les observations (pour un modèle idéal, les points seraient tous sur la bissectrice), et les résidus normalisés sur la forme d'un diagramme en bâtons. Ce dernier graphique permet de rapidement voir si un nombre anormal de données sort de l'intervalle]-2, 2[sachant que ce dernier, sous hypothèse de normalité, doit contenir environ 95% des données.

Si vous avez sélectionné des données à utiliser pour calculer des **prédictions sur de nouvelles observations**, le tableau correspondant est ensuite affiché.

Exemple

Un exemple d'utilisation de la méthode des doubles moindres carrés est disponible sur le Centre d'aide XLSTAT à l'adresse

http://www.xlstat.com/demo-slsf.htm

Bibliographie

Akaike H. (1973). Information theory and the extension of the maximum likelihood principle. In: Second International Symposium on Information Theory. (Eds: V.N. Petrov and F. Csaki). Academiai Kiadó, Budapest. 267-281.

Amemiya T. (1980). Selection of regressors. *International Economic Review*, **21**, 331-354.

Mallows C.L. (1973). Some comments on Cp. *Technometrics*, **15**, 661-675.

Theil, H. (1953a). Repeated least square applied to complete equation systems. mimeo, Central Planning Bureau, The Hague.

Theil, H. (1953b), Estimation and simultaneous correlation in complete equation systems. Central Planning Bureau, The Hague.

Régression PLS/PCR

Utilisez ce module pour modéliser et prédire les valeurs d'une ou plusieurs variables quantitatives en fonction d'une combinaison linéaire d'une ou plusieurs variables explicatives quantitatives et/ou qualitatives en vous affranchissant des contraintes de la régression linéaire pour ce qui concerne la distribution des variables et le nombre de variables que l'on peut inclure.

Dans cette section :

Description

Les trois méthodes de régression auxquelles ce module donne accès ont pour propriétés communes de générer des modèles mettant en jeu des combinaisons linéaires de variables explicatives. La différence entre les trois méthodes vient essentiellement de la façon dont sont traitées les structures de corrélations entre les variables.

Régression PCR

La régression PCR (*Principal Components Regression*) ou régression sur composantes principales comprend trois étapes : on réalise d'abord une ACP (Analyse en Composantes Principales) sur le tableau des variables explicatives, puis on effectue une régression OLS sur les composantes retenues, puis on calcule les paramètres du modèle sur les variables d'origine.

L'ACP permet de passer d'un tableau X comprenant n observations décrites par p variables à un tableau S de n observations décrites par q composantes, où q est inférieur ou égal à p et tel que ($S'S$) est inversible. Une sélection supplémentaire peut être effectuée de telle sorte que seuls les r composantes les plus corrélées avec la variable Y soient gardées pour la régression OLS. On obtient alors le tableau R.

Le calcul de la régression OLS s'effectue sur le tableau R. On obtient alors les paramètres correspondant à chacun des r facteurs. Afin de palier le problème d'interprétation des paramètres ainsi obtenus, XLSTAT effectue automatiquement les calculs nécessaires pour obtenir les paramètres et les intervalles de confiance pour les variables de départ.

Régression PLS

Cette méthode est rapide, efficace et optimale pour un critère de minimisation des covariances bien maîtrisé. Son utilisation est recommandée dans le cas où un grand nombre de variables explicatives est utilisé, ou lorsqu'il y a de fortes colinéarités entre les variables.

L'idée de la régression PLS (*Partial Least Squares*) est de créer à partir d'un tableau de n observations décrites par p variables, un ensemble de h composantes avec $h < p$. La méthode de construction des composantes diffère de celle de l'ACP, et présente l'avantage de bien s'accommoder de la présence de données manquantes. La détermination du nombre de composantes à retenir est en général fondée sur un critère mettant en jeu une validation croisée. L'utilisateur peut aussi fixer lui-même le nombre de composantes à retenir.

On distingue souvent la PLS1 de la méthode PLS2. La PLS1 concerne le cas où il y a une seule variable dépendante, la PLS2 celui où il y a plusieurs variables dépendantes. Les algorithmes utilisés dans XLSTAT sont tels que la PLS1 est un cas particulier de la PLS2. La distinction ne sera donc pas faite ici.

Dans le cas des méthodes OLS et PCR, si l'on doit calculer les modèles pour plusieurs variables dépendantes, le calcul des modèles consiste en une simple boucle sur les colonnes du tableau des variables dépendantes. Dans le cas de la régression PLS, la structure de covariance du tableau des variables dépendantes influe aussi sur les calculs.

L'équation du modèle de la régression PLS avec h composantes est donnée par

$$\begin{aligned} Y &= T_h C_h' + E_h \\ &= X W_k^* C_h' + E_h \\ &= X W_h (P_h' W_h)^{-1} C_h' + E_h \end{aligned}$$

où Y est la matrice des variables dépendantes, X celle des variables explicatives, et où T_h, C_h, W_h^* , W_h et P_h, sont des matrices générées par l'algorithme PLS, et où E_h est la matrice des résidus.

La matrice B des coefficients de régression de Y sur X en utilisant h composantes générées par l'algorithme de régression PLS est donc définie par

$$B = W_h (P_h' W_h)^{-1} C_h'$$

Remarque : il s'agit donc comme en régression OLS ou PCR d'un modèle linéaire.

Remarques :

Les trois méthodes donnent le même résultat si le nombre de composantes issues de l'ACP (en régression PCR) ou de la PLS (régression PLS) est égal au nombre de variables explicatives sélectionnées.

En régression PLS, les composantes sont créées de fait de telle sorte qu'elles expliquent au mieux Y, alors qu'en PCR elles sont au départ créées uniquement en fonction de X. XLSTAT

permet de corriger partiellement ce désavantage en proposant de sélectionner les composantes les plus corrélées avec Y.

Analyse discriminante PLS

L'analyse discriminante PLS (PLS-DA) permet d'utiliser la méthode PLS pour expliquer et prédire l'appartenance d'individus à plusieurs classes, sur la base de variables explicatives quantitatives ou qualitatives. Pour cela, XLSTAT-PLS utilise l'algorithme PLS2 en prenant comme variable dépendante le tableau disjonctif complet obtenu à partir de la variable de classification qualitative.

L'analyse discriminante PLS permet d'effectuer une discrimination en utilisant les propriétés de l'algorithme PLS. Elle peut ainsi s'appliquer au cas de jeux de données avec peu d'observations et beaucoup de variables explicatives. Les composantes PLS sont utilisées et les mêmes options que dans le cas de la régression PLS sont disponibles. Elle permet aussi de n'utiliser que les données disponibles dans le cas de données manquantes et s'adapte très bien au cas de multicolinéarité entre les variables explicatives.

On obtient donc autant de modèles que de modalités de la variable dépendante. Une observation est associée à la classe pour laquelle la valeur de l'équation du modèle est maximale.

Soit une variable dépendante à K modalités a_1, …, a_K. Pour chaque modalité de la variable dépendante, on peut calculer pour chaque observation l'équation :

$$F(y_i, a_k) = b_0 + \sum_{j=1}^{p} b_i x_{ij}$$

Avec a_k une des modalités, b_0 constante du modèle associé à la modalité a_k obtenu par régression PLS, p nombre de variables indépendantes et b_i coefficients de ce même modèle.

L'observation i est classée dans la classe k si :

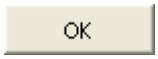
$$k^* = \arg\max_k F(y_i, a_k)$$

L'analyse discriminante PLS offre une alternative intéressante à l'analyse discriminante linéaire classique.

Des sorties équivalentes à l'analyse discriminante de XLSTAT sont ainsi disponibles dans l'outil PLS : la matrice de confusion, les valeurs prédites ainsi que la validation et la prédiction.

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.


: cliquez sur ce bouton pour changer la façon dont vous chargez les données dans XLSTAT. Si le bouton possède une icône de souris, XLSTAT vous permet de faire une sélection à la souris des données. Si le bouton possède une icône de liste, XLSTAT vous permet de sélectionner les données à partir d'une liste. Si le bouton possède une icône orange, des boutons complémentaires avec un point d'interrogation  sont affichés pour vous permettre d'importer les données à partir de fichiers.

Onglet **Général** :

Y / Variables dépendantes :

Quantitatives/Qualitatives : sélectionnez la ou les variables dépendantes quantitatives (qualitatives dans le cas de la PLS-DA). Les données sélectionnées doivent être de type numérique (les données peuvent être nominales dans le cas de la PLS-DA). Si le libellé des variables a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

X / Variables explicatives :

Quantitatives : activez cette option si vous voulez inclure une ou plusieurs variables explicatives quantitatives. Sélectionnez alors la ou les variables correspondantes sur la feuille Excel. Les données sélectionnées doivent être de type numérique. Si le libellé des variables a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

Qualitatives : activez cette option si vous voulez inclure une ou plusieurs variables explicatives qualitatives dans le modèle. Sélectionnez alors la ou les variables correspondantes sur la feuille Excel. Les données sélectionnées peuvent être de tout type, mais les données numériques sont automatiquement considérées comme nominales. Si le libellé des variables a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

Méthode : choisissez la méthode de régression à utiliser :

- **PLS-R** : activez cette option pour calculer une régression avec la méthode des moindres carrés partiels (Partial Least Squares).
- **PLS-DA** : activez cette option pour calculer une analyse discriminante avec la méthode des moindres carrés partiels (Partial Least Squares).
- **PCR** : activez cette option pour calculer une régression sur les composantes principales (Principal Components Regression).

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des variables : activez cette option si la première ligne des données sélectionnées (variables dépendantes, explicatives, poids et libellés des observations) contient un libellé.

Libellés des observations : activez cette option si vous voulez utiliser des libellés d'observations disponibles sur une feuille Excel pour l'affichage des résultats. Si l'option « Libellés des variables » est activée, la première cellule de la sélection doit comprendre un en-tête. Si vous n'activez pas cette option, des libellés seront automatiquement créés (Obs1, Obs2, …).

Poids des observations : activez cette option si vous voulez pondérer les observations. Si vous n'activez pas cette option, les poids seront tous considérés comme valant 1. Les poids doivent être impérativement supérieurs ou égaux à 0. Un poids de 2 est équivalent à répéter deux fois la même observation. Si un en-tête de colonne a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

Poids dans la régression : cette option n'est active que pour les régressions PCR et OLS. Activez cette option si vous voulez effectuer une régression par les moindres carrés pondérés. Si vous n'activez pas cette option, les poids seront tous considérés comme valant 1. Les poids doivent être impérativement supérieurs ou égaux à 0. Si un en-tête de colonne a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

Onglet **Options** :

Options communes :

Intervalle de confiance (%) : entrez l'étendue en pourcentage de l'intervalle de confiance à utiliser pour les différents tests, et pour le calcul des intervalles de confiance autour des paramètres et des prédictions. Valeur par défaut : 95.

Interactions / Niveau : activez cette option pour inclure des interactions dans le modèle puis entrez le niveau maximum d'interaction (valeur comprise entre 1 et 4).

Options pour la régression PLS :

Conditions d'arrêt :

- **Automatique** : activez cette option pour que XLSTAT détermine automatiquement le nombre de composantes à retenir.
- **Nombre fixé** : activez cette option pour fixer le nombre de composantes à prendre en compte dans le modèle. La valeur par défaut est 2.
- **Seuil Qi²** : activez cette option pour fixer la valeur seuil du critère Qi² utilisée pour déterminer si l'apport d'une composante est significatif ou non pour l'une des variables dépendantes. La valeur par défaut est 0.0975 et correspond à 1-0.95².
- **Seuil Qi² (global)** : activez cette option pour fixer la valeur seuil du critère Qi² utilisée pour déterminer si l'apport d'une composante est significatif ou non pour l'ensemble des variables dépendantes. La valeur par défaut est 0.0975 et correspond à 1-0.95².
- **Amélioration du Qi²** : activez cette option pour fixer la valeur seuil du critère d'amélioration du Qi^2, noté Qi^2Imp utilisée pour déterminer si l'apport d'une composante est significatif ou non. La valeur par défaut est 0.05 et correspond à 5% d'amélioration. La valeur de ce critère est donnée par

$$Q^2(h)\,\mathrm{Imp} = \frac{Q^2(h) - Q^2(h-1)}{Q^2(h-1)}$$

- **Press minimum** : activez cette option pour que le nombre de composantes retenues corresponde au modèle donnant le coefficient de Press minimal.

X / Variables explicatives : * **Centrer** : activez cette option si vous voulez centrer les variables explicatives avant de commencer les calculs.

- **Réduire** : activez cette option si vous voulez réduire les variables explicatives avant de commencer les calculs.

Algorithme : la différence entre les deux approches peut être seulement vu si l'option jackknife est sélectionnée pour la validation croisée et si les données sont centrées ou normalisées.

- **Rapide** : activez cette option pour utiliser un algorithme plus rapide. L'algorithme évite le recentrage ou la renormalisation des ensembles d'entraînement de jackknife.
- **Précis** : activez cette option pour utiliser un algorithme plus lent mais néanmoins plus précis. L'algorithme recentre ou renormalise les ensembles d'entraînement de jackknife.

Validation croisée :

- **Aucune** : activez cette option pour ne pas effectuer de validation croisée. Dans ce cas les calculs seront plus rapides, mais les Q² et les intervalles de confiance ne pourront pas

être calculés.

- **Jackknife (LOO)** : activez cette option pour effectuer une validation croisée basée sur un jackknife de type « leave one out ». Cette méthode n'est pas utilisable au-delà de 100 observations en raison de son coût en mémoire.

- **Jackknife** : activez cette option pour effectuer une validation croisée basée sur un jackknife avec répartition des données en k groupes. Un maximum de 100 groupes est accepté.

Options pour la régression PCR :

ACP normée : activez cette option pour effectuer une ACP sur la matrice de corrélation. Désactivez cette option pour effectuer une ACP sur la matrice de covariance.

Filtrer les composantes : vous pouvez activer l'une ou les deux options suivantes afin de réduire le nombre de composantes utilisées dans le modèle :

- **% minimum** : activez cette option puis saisissez le pourcentage minimum de la variabilité totale que doivent représenter les composantes sélectionnées.

- **Nombre maximum** : activez cette option pour fixer le nombre maximum de composantes à prendre en compte.

Trier les composantes : choisissez l'une des options suivantes afin de déterminer quel critère est utilisé pour trier les composantes avant que soient pris en compte les critères « % minimum » ou « Nombre maximum » :

- **Corrélations avec les Y** : activez cette option pour que la sélection des composantes se fasse après un tri décroissant suivant le carré du coefficient de corrélation (R^2) entre la variable Y et les composantes. Cette option est recommandée.

- **Valeurs propres** : activez cette option pour que la sélection des composantes se fasse après un tri décroissant suivant les valeurs propres associées aux composantes.

Constante fixée : activez cette option pour fixer la constante du modèle de régression à une valeur que vous devez ensuite saisir (0 par défaut).

Tolérance : activez cette option pour permettre à l'algorithme de calcul de la régression OLS ne pas prendre en compte les variables qui seraient soit constantes soit trop corrélées avec d'autres variables déjà utilisées dans le modèle (0.0001 par défaut).

Onglet **Validation** :

Validation : activez cette option si vous souhaitez utiliser une partie des données sélectionnées pour valider le modèle.

Jeu de validation : choisissez l'une des options pour définir le mode de sélection des observations utilisées pour la validation :

- **Aléatoire** : les observations sont sélectionnées de manière aléatoire. Le « Nombre d'observations » doit alors être saisi.
- **N dernières lignes** : les N dernières observations sont sélectionnées pour la validation. Le « Nombre d'observations » N doit alors être saisi.
- **N premières lignes** : les N premières observations sont sélectionnées pour la validation. Le « Nombre d'observations » N doit alors être saisi.
- **Variable de groupe** : si vous choisissez cette option, vous devez ensuite sélectionner une variable indicatrice composée de 0 pour les observations à utiliser pour le calcul du modèle, et de 1 pour les observations à utiliser pour la validation du modèle.

Onglet **Prédiction** :

Prédiction : activez cette option si vous souhaitez sélectionner des données à utiliser en mode prédiction. Si vous activez cette option, vous devez veiller à ce que les données de prédiction soient organisées comme les données d'estimation : mêmes variables, même ordre dans les sélections. En revanche vous ne devez pas sélectionner de libellés de variables : la première ligne des sélections décrites ci-dessous doit être une ligne de données.

Quantitatives : activez cette option pour sélectionner la ou les variables quantitatives explicatives. La première ligne ne doit pas comprendre d'en- tête.

Qualitatives : activez cette option pour sélectionner la ou les variables qualitatives explicatives. La première ligne ne doit pas comprendre d'en-tête.

Libellés des observations : activez cette option si vous voulez utiliser des libellés d'observations disponibles sur une feuille Excel pour l'affichage des résultats. La première ligne ne doit pas comprendre d'en-tête. Si vous n'activez pas cette option, des libellés seront automatiquement créés (PredObs1, PredObs2, …).

Libellés des variables : activez cette option si la première ligne des données sélectionnées (variables explicatives, libellés des observations) contient un libellé.

Onglet **Données manquantes** :

Ces options ne sont disponibles que dans le cas des régressions PCR et OLS. Pour la régression PLS, la gestion des données manquantes fait partie de la méthode.

Supprimer les observations : activez cette option pour supprimer les observations comportant des données manquantes.

- **Vérifier pour chaque Y séparément** : choisissez cette option si vous voulez que lorsque, pour une observation donnée, il y a des données manquantes uniquement dans les Y, l'observation ne soit supprimée que si la donnée correspondant au Y en cours de modélisation est manquante.

- **Pour tous les Y** : choisissez cette option pour supprimer toutes les observations pour lesquelles des Y sont manquants.

- Remarque : les deux alternatives ci-dessus sont sans effet s'il n'y a qu'un seul Y.

Ignorer les données manquantes : si vous choisissez cette option, pour les données manquantes correspondant aux variables dépendantes XLSTAT essaiera de les estimer à partir du modèle obtenu. Pour celles correspondant aux variables explicatives, les observations correspondantes seront conservées dans la mesure du possible pour estimer la matrice de variance covariance (suppression par paire).

Estimer les données manquantes : activez cette option pour estimer les données manquantes avant le début des calculs.

- **Moyenne ou mode** : activez cette option pour estimer les données manquantes en utilisant la moyenne (variables quantitatives) ou le mode (variables qualitatives) pour les variables correspondantes.

- **Plus proche voisin** : activez cette option pour estimer les données manquantes d'une observation en recherchant le plus proche voisin de l'observation.

Onglet **Sorties** :

Options communes :

Statistiques descriptives : activez cette option pour afficher les statistiques descriptives pour l'ensemble des variables sélectionnées.

Corrélations : activez cette option pour afficher la matrice de corrélation pour les variables quantitatives (dépendantes et explicatives).

Coefficients normalisés : activez cette option pour afficher les paramètres normalisés du modèle (coefficients bêta).

Equation : activez cette option pour afficher explicitement l'équation du modèle.

Prédictions et résidus : activez cette option pour afficher les prédictions et les résidus pour l'ensemble des observations.

Options pour la régression PLS-R :

Composantes t, u et u~ : activez cette option pour afficher les tableaux des composantes. Si cette option n'est pas activée, les graphiques correspondants ne sont pas affichés.

Vecteurs c, w, w* et p : activez cette option pour afficher les tableaux des vecteurs générés par l'algorithme PLS. Si cette option n'est pas activée, les graphiques correspondants ne sont pas affichés.

VIP : activez cette option pour afficher le tableau et les graphiques correspondant aux *Variable Importance for the Projection* (VIP).

Intervalles de confiance : activez cette option pour calculer et afficher les intervalles de confiance autour des « coefficients standardisés ». Les calculs utilisent une méthode Jackknife.

Détection des valeurs extrêmes : activez cette option pour afficher le tableau et les graphiques des valeurs extrêmes.

Options pour la régression PLS-DA :

Composantes t, u et u~ : activez cette option pour afficher les tableaux des composantes. Si cette option n'est pas activée, les graphiques correspondants ne sont pas affichés.

Vecteurs c, w, w* et p : activez cette option pour afficher les tableaux des vecteurs générés par l'algorithme PLS. Si cette option n'est pas activée, les graphiques correspondants ne sont pas affichés.

VIP : activez cette option pour afficher le tableau et les graphiques correspondant aux *Variable Importance for the Projection* (VIP).

Intervalles de confiance : activez cette option pour calculer et afficher les intervalles de confiance autour des « coefficients standardisés ». Les calculs utilisent une méthode Jackknife.

Détection des valeurs extrêmes : activez cette option pour afficher le tableau et les graphiques des valeurs extrêmes.

Matrice de confusion : activez cette option pour afficher le tableau permettant de visualiser les nombres d'observations bien et mal classées pour chacune des classes.

Options pour la régression PLS-R :

Coordonnées des variables : activez cette option pour afficher les coordonnées des variables (*factor loadings* en anglais). Les coordonnées sont égales aux corrélations entre les composantes principales et les variables d'origine dans le cas d'une ACP normée.

Corrélations Composantes/Variables : activez cette option pour afficher les corrélations entre les composantes principales et les variables d'origine.

Coordonnées des observations : activez cette option pour afficher les coordonnées des observations (*factor scores* en anglais) dans le nouvel espace créé par l'ACP. Ces coordonnées sont ensuite utilisées dans l'étape OLS de la régression PCR.

Options pour les régressions PCR :

Analyse de la variance : activez cette option pour afficher le tableau d'analyse de la variance.

Prédictions ajustées : activez cette option pour calculer et afficher les prédictions ajustées dans le tableau des prédictions et résidus.

- **Diagnostics d'influence** : activez cette option pour calculer et afficher le tableau des statistiques permettant d'identifier les observations ayant une influence sur les prédictions ou sur les coefficients associés à certaines variables explicatives (voir section résultats).

Onglet **Graphiques** :

Options communes :

Graphiques de régression : activez cette option pour afficher les graphiques de régression :

- **Coefficients normalisés** : activez cette option pour afficher sur un graphique les paramètres normalisés du modèle avec leur intervalle de confiance.

- **Prédictions et résidus** : activez cette option pour afficher les graphiques suivants :

(1) Droite de régression : ce graphique n'est affiché que s'il n'y a qu'une seule variable explicative, et que cette variable est quantitative.

(2) Variable explicative versus résidus normalisés : ce graphique n'est affiché que s'il n'y a qu'une seule variable explicative, et que cette variable est quantitative.

(3) Variable dépendante versus résidus normalisés.

(4) Prédictions pour la variable dépendante versus variable dépendante.

(5) Graphique en bâtons des résidus normalisés.

- **Intervalles de confiance** : activez cette option pour afficher les intervalles de confiance sur les graphiques (1) et (4).

Options pour les régressions PLS et PCR :

Graphiques de corrélations : activez cette option pour afficher les graphiques mettant en jeu des corrélations entre des composantes et des variables initiales. Dans le cas de la PCR, activez cette option pour afficher le cercle des corrélations.

- **Vecteurs** : activez cette option pour afficher les variables d'origine sous forme de vecteurs.

Graphiques des observations : activez cette option pour afficher les graphiques de représentation des observations dans le nouvel espace.

- **Etiquettes** : activez cette option pour afficher les étiquettes des observations sur les graphiques. Le nombre d'étiquettes affichées peut être modulé à l'aide de l'option de filtrage.

Biplots : activez cette option pour afficher les graphiques de représentation simultanée des observations et des variables d'origine dans le nouvel espace.

- **Vecteurs** : activez cette option pour afficher les variables d'origine sous forme de vecteurs.

- **Etiquettes** : activez cette option pour afficher les étiquettes des observations sur les biplots. Le nombre d'étiquettes affichées peut être modulé à l'aide de l'option de filtrage.

Etiquettes colorées : activez cette option pour afficher les étiquettes de variables et d'observations de la même couleur que les points correspondants

Filtrer : activez cette option pour fixer le nombre de points affichés sur les graphiques :

- **Aléatoire** : les observations sont sélectionnées de manière aléatoire. Le « Nombre d'observations » doit alors être saisi.
- **N premières lignes** : les N premières observations sont affichées. Le « Nombre d'observations » N doit alors être saisi.
- **N dernières lignes** : les N dernières observations sont affichées. Le « Nombre d'observations » N doit alors être saisi.
- **Variable de groupe** : si vous choisissez cette option, vous devez ensuite sélectionner une variable indicatrice composée de 1 pour les observations à afficher, et de 1 pour les observations à ne pas afficher.

Résultats

Statistiques descriptives : les tableaux de statistiques descriptives présentent pour toutes les variables sélectionnées des statistiques simples. Pour les variables dépendantes (en bleu) et les variables explicatives quantitatives, sont affichés le nombre d'observations, le nombre de données manquantes, le nombre de données non manquantes, la moyenne, et l'écart-type (non biaisé). Pour les variables explicatives qualitatives sont affichés le nom des différentes modalités ainsi que leur fréquence respective.

Matrice de corrélation : ce tableau est affiché afin de vous permettre d'avoir un aperçu des corrélations entre les différentes variables sélectionnées.

Résultats de la régression PLS :

Le premier tableau présente des indices de **qualité du modèle** sous forme de contribution cumulée des composantes aux indices :

- L'indice **Q²cum** est une mesure de l'apport global des h premières composantes à la qualité prédictive du modèle (et de ses sous-modèles s'il y a plusieurs variables dépendantes). L'indice Q²cum(h) est défini par :

$$Q^2cum(h) = 1 - \prod_{j=1}^{h} \frac{\sum_{k=1}^{q} PRESS_{kj}}{\sum_{k=1}^{q} SCE_{k(j-1)}}$$

Cet indice fait intervenir un rapport des coefficients de PRESS (impliquant donc une validation croisée) et de la somme des carrés des erreurs (SCE) pour un modèle utilisant une composante de moins. La recherche du maximum de Q²cum revient donc à chercher le modèle le plus stable possible.

- L'indice **R²Ycum** est la somme des coefficients de détermination entre les variables dépendantes et les h premières composantes. C'est donc une mesure du pouvoir explicatif des h premières composantes pour les variables dépendantes du modèle.

- L'indice **R²Xcum** est la somme des coefficients de détermination entre les variables explicatives et les h premières composantes. C'est donc une mesure du pouvoir explicatif des h premières composantes pour les variables explicatives du modèle.

Un diagramme en bâtons est ensuite affiché afin de permettre une visualisation de l'évolution des trois indices en fonction du nombre de composantes. Si les R²Ycum et R²Xcum croissent nécessairement avec le nombre de composantes, ce n'est pas le cas pour Q²cum.

Le tableau suivant correspond à la **matrice de corrélation** des variables explicatives et dépendantes avec les composantes t et u. Un graphique permet ensuite de visualiser les corrélations avec les composantes t.

Le tableau des **vecteurs w** est ensuite affiché, suivi des tableaux des **vecteurs w*** et des **vecteurs c** qui, comme il est montré dans la section « Description », interviennent directement dans le modèle. Si à $h = 2$ correspond un modèle acceptable, il est démontré que la projection des vecteurs X sur les vecteurs Y sur le **graphique des variables sur les w*/c**, fournit une idée d'une part du signe dans le modèle des coefficients correspondant, d'autre part du poids relatif des variables de départ pour l'explication des variables dépendantes.

Le tableau suivant correspond aux **coordonnées des observations** dans l'espace des composantes **t**. Le graphique est ensuite affiché. Si des observations de validation ont été sélectionnées, elles sont affichées sur ce graphique.

Le tableau des **coordonnées normalisées** est ensuite affiché. Ces coordonnées sont égales à la corrélation entre les variables indicatrices des observations et les composantes t. Ces coordonnées sont utilisées pour le graphique des **corrélations** qui suit, et qui permet la visualisation simultanée des observations, des variables dépendantes et des variables explicatives. Un exemple d'interprétation de ce graphique est disponible dans Tenenhaus (2003).

Le tableau suivant correspond aux **coordonnées des observations** dans l'espace des composantes **u** puis dans celui des composantes **u~**. Le graphique est ensuite affiché. Si des observations de validation ont été sélectionnées elles sont affichées sur ce graphique.

Le tableau des **indices de qualité Q²** permet de voir comment les composantes contribuent à l'explication des variables dépendantes. Le tableau des **indices de qualité Q² cumulé** permet de mesurer la qualité associée à un espace de dimension croissante.

Les tableaux des **R² et des redondances** entre les variables de départ (dépendantes et explicatives) et les composantes t et u permettent de mesurer le pouvoir explicatif des composantes t et u~ tant au sens du R^2 qu'au sens de la redondance. La redondance entre un tableau X (n lignes et p variables) et une composante c est la part de la variance de X expliquée par c. On la définit comme la moyenne des carrés des coefficients de corrélation entre les variables et la composante :

$$Rd(X,c) = \frac{1}{p}\sum_{j=1}^{p} R^2(x_j,c)$$

Des redondances on peut alors déduire les VIP (**Variable Importance for the Projection**) qui mesurent l'importance d'une variable explicative pour la construction des composantes t. La VIP pour la variable explicative j et la composante h est définie par

$$VIP_{hj} = \sqrt{\frac{p}{\sum_{i=1}^{h} Rd(Y,t_i)} \sum_{i=1}^{h} {w_{ij}}^2 Rd(Y,t_i)}$$

Sur les graphiques des VIP (un diagramme en bâton par composante), une limite est tracée pour identifier les VIP supérieures à 0.8 ; il s'agit d'un seuil empirique proposé par Wold (1995) permettant d'identifier les variables fortement contributrices au modèle.

Le dernier suivant permet la **détection des valeurs extrêmes**. Les DModX (distances au modèle des observations dans l'espace des X) permettent d'identifier les valeurs anormales des variables explicatives, tandis que les DModY (distances au modèle des observations dans l'espace des Y) permettent d'identifier les valeurs anormales des variables dépendantes. Sur les graphiques correspondants sont affichés les seuils DCrit à partir desquels on peut considérer qu'une valeur de DMod est anormalement élevée. Les DCrit sont calculés en utilisant les valeurs seuil, classiquement calculées pour les box plots. La valeur de DModX pour la i-ème observation est définie par :

$$DModX_i = \sqrt{\frac{n}{n-h-1} \frac{\sum_{j=1}^{p} e(X,t)^2{}_{ij}}{p-h}}$$

où les $e(X,t)_{ij},\ (i=1,\ldots,n)$ sont les résidus de la régression de X sur la j-ème composante. La valeur de DModY pour la i-ième observation est définie par :

$$DModY_i = \sqrt{\frac{\sum_{j=1}^{q} e(Y,t)_{ij}^2}{q-h}}$$

où q est le nombre de variables dépendantes, et où les $e(Y,t)_{ij},\ (i=1,\ldots,n)$ sont les résidus de la régression de Y sur la j-ème composante.

Le tableau qui suit présente les paramètres des modèles pour les différentes variables dépendantes, suivi des équations correspondantes si le nombre de variables explicatives est inférieur à 20.

Pour chacune des variables dépendantes est ensuite affichée une série de tableaux et graphiques.

Statistiques d'ajustement : dans ce tableau sont affichées les statistiques relatives à l'ajustement du modèle de régression, dont les définitions sont données dans la section consacrée

Le tableau des **coefficients normalisés** (aussi appelés coefficients bêta) permet de comparer le poids relatif des variables dans le modèle. Pour le calcul des intervalles de confiance, dans le cadre de la PLS, les formules basées sur les hypothèses de normalité utilisées en régression OLS ne sont plus valables. Une méthode bootstrap proposée par Tenenhaus *et al* (2004) permet d'estimer les intervalles de confiance. Plus la valeur absolue d'un coefficient est élevée, plus le poids de la variable correspondante est important. Lorsque l'intervalle de confiance autour des coefficients normalisés comprend la valeur 0 (cela est facilement visible sur le graphique des coefficients normalisés), le poids d'une variable dans le modèle n'est pas significatif.

Dans le tableau des **prédictions et résidus** sont donnés pour chaque observation la valeur observée de la variable dépendante, la prédiction du modèle, les résidus et les intervalles de confiance. Deux types d'intervalles de confiance sont affichés : un intervalle de confiance autour de la moyenne (correspondant au cas où l'on ferait la prédiction pour un nombre infini d'observations avec un ensemble de valeurs données des variables explicatives) et un intervalle autour de la prédiction ponctuelle (correspondant au cas d'une prédiction isolée pour des valeurs données des variables explicatives). Le second intervalle est toujours plus grand que le premier, les aléas étant plus importants. Si des données de validation ont été sélectionnées, elles sont affichées en fin de tableau.

Les **trois graphiques** affichés ensuite permettent de visualiser respectivement

- l'évolution des résidus en fonction de la variable dépendante,
- la distance entre les valeurs prédites et observées (pour un modèle idéal, les points seraient tous sur la bissectrice),
- le diagramme en bâtons des résidus.

Si vous avez sélectionné des données à utiliser pour calculer des prédictions, le tableau des prédictions est ensuite affiché.

Résultats spécifiques à l'analyse discriminante PLS :

Fonctions de classement : les fonctions de classement peuvent être utilisées pour déterminer à quelle classe doit être affectée une observation sur la base des valeurs prises pour les différentes variables explicatives. Une observation est affectée à la classe pour laquelle la fonction de classement est la plus élevée.

Classification a priori, a posteriori et scores de classement : dans ce tableau sont affichés pour chaque observation, sa classe d'appartenance définie par la variable dépendante, la classe d'appartenance telle que déduite des fonctions de classement et les valeurs des fonctions de classement associées à chacune des classes.

Matrice de confusion pour l'échantillon d'estimation : En utilisant les classifications a priori et a posteriori, on déduit la matrice de confusion, ainsi que le pourcentage global d'observations bien classées.

Résultats de la régression PCR :

La régression PCR requérant le calcul d'une Analyse en Composantes Principales, les résultats concernant cette dernière sont affichés.

Valeurs propres : les valeurs propres et le graphique (*scree plot*) correspondant sont affichés. Le nombre de valeurs propres est égal au nombre de valeurs propres non nulles. Si un filtrage a été demandé, il est appliqué au niveau de la régression elle-même.

Si les options de sorties correspondantes ont été activées, XLSTAT affiche ensuite les **coordonnées des variables** dans le nouvel espace, puis les corrélations entre les variables d'origine et les composantes dans le nouvel espace. Les **corrélations** sont égales aux coordonnées des variables dans le cas d'une ACP normée. Les **coordonnées des observations** dans le nouvel espace sont affichées dans un troisième tableau, et constituent les données utilisées ensuite pour la régression. Si des données de validation ont été sélectionnées, elles sont affichées en fin de tableau. Si l'option correspondante a été activée les biplots sont affichés.

Si l'option de filtrage des composantes, s'appuyant sur les corrélations avec les variables dépendantes a été choisie, les composantes retenues pour la régression sont celles présentant les plus forts coefficients de détermination (R^2) avec les variables dépendantes. La matrice des coefficients de **corrélation entre les composantes et les variables dépendantes** est alors affichée. Le nombre de composantes retenues dépend du nombre de valeurs propres et des options choisies ("% minimum" ou "Max composantes").

Si l'option de filtrage des composantes s'appuyant sur les valeurs propres a été choisie, les composantes retenues pour la régression sont celles présentant les plus fortes valeurs propres. Le nombre de composantes retenues dépend du nombre de valeurs propres et des options choisies ("% minimum" ou "Max composantes").

Coefficients d'ajustement : dans ce tableau sont affichées les statistiques relatives à l'ajustement du modèle de régression :

- **Observations** : le nombre d'observations prises en compte dans les calculs. Dans les formules présentées ci-dessous n désigne le nombre d'observations.
- **Somme des poids** : la somme des poids des observations prises en compte dans les calculs. Dans les formules présentées ci-dessous W désigne la somme des poids.
- **DDL** : le nombre de degrés de liberté pour le modèle retenu (correspondant à la partie erreurs).
- **R²** : le coefficient de détermination du modèle. La valeur de ce coefficient est comprise entre 0 et 1. XLSTAT le calcule comme suit :

$$R^2 = \frac{\sum_{i=1}^{n} w_i(\hat{y}_i - \bar{y}_i)^2}{\sum_{i=1}^{n} w_i(y_i - \bar{y}_i)^2} \text{ avec } \bar{y} = \frac{1}{W}\sum_{i=1}^{n} w_i y_i$$

Le R^2 s'interprète comme la proportion de la variabilité de la variable dépendante expliquée par le modèle. Plus le R^2 est proche de 1, meilleur est le modèle. L'inconvénient du R^2 est qu'il ne prend pas en compte le nombre de variables utilisées pour ajuster le modèle.

- **R² ajusté** : le coefficient de détermination ajusté du modèle. Le R^2 ajusté peut être négatif si le R^2 est voisin de zéro. Ce coefficient n'est affiché que si la constante du modèle n'est pas fixée par l'utilisateur. Sa valeur est définie par

$$\hat{R}^2 = 1 - (1 - R^2)\frac{W - 1}{W - p - 1}$$

Le R^2 ajusté est une correction du R^2 qui permet de prendre en compte le nombre de variables utilisées dans le modèle.

- **MCE** : la moyenne des carrés des erreurs (MCE) est définie par

$$MCE = \frac{1}{W - p^*}\sum_{i=1}^{n} w_i(y_i - \hat{y}_i)^2$$

- **RMCE** : la racine de la moyenne des carrés des erreurs (RMCE) est la racine carrée de la MCE.

- **MAPE** : la *Mean Absolute Percentage Error* est calculée comme suit :

$$MAPE = \frac{100}{W}\sum_{i=1}^{n} w_i \left|\frac{y_i - \hat{y}_i}{y_i}\right|$$

- **DW** : le coefficient de Durbin-Watson est défini par

$$DW = \frac{\sum_{i=2}^{n}[(y_i - \hat{y}_i) - (y_{i-1} - \hat{y}_{i-1})]^2}{\sum_{i=1}^{n} w_i(y_i - \hat{y}_i)^2}$$

Ce coefficient correspond au coefficient d'autocorrélation d'ordre 1 et permet de vérifier que les résidus du modèle ne sont pas autocorrélés, sachant que l'indépendance des résidus est l'une des hypothèses de base de la régression linéaire. L'utilisateur pourra se référer à une table des coefficients de Durbin-Watson pour vérifier si l'hypothèse d'indépendance des résidus est acceptable.

- **Cp** : le coefficient Cp de Mallows est défini par

$$Cp = \frac{SCE}{\hat{\sigma}} + 2p^* - W$$

où SCE est la somme du carré des erreurs pour le modèle avec p variables explicatives, et où $\hat{\sigma}$ correspond à l'estimateur de la variance des résidus pour le modèle comprenant toutes les variables explicatives. Plus le coefficient Cp est proche de p* moins le modèle est biaisé.

- **AIC** : le critère d'information d'Akaike (Akaike's Information Criterion) est défini par

$$AIC = W\ln(\frac{SCE}{W}) + 2p^*$$

Ce critère proposé par Akaike (1973) dérive de la théorie de l'information, et s'appuie sur la mesure de Kullback et Leibler (1951). C'est un critère de sélection de modèles qui pénalise les modèles pour lesquels l'ajout de nouvelles variables explicatives n'apporte pas suffisamment d'information au modèle, l'information étant mesurée au travers de la SCE. On cherche à minimiser le critère AIC.

- **SBC** : le critère bayésien de Schwarz (Schwarz's Bayesian Criterion) est défini par

$$SBC = W \ln(\frac{SCE}{W}) + \ln(W)p^*$$

Ce critère proposé par Schwarz (1978) est proche du critère AIC, et comme ce dernier on cherche à le minimiser.

- **PC** : le critère de prédiction d'Amemiya (Amemiya's Prediction Criterion) est défini par

$$PC = \frac{(1 - R^2)(W + p^*)}{W - p^*}$$

Ce critère proposé par Amemiya (1980) permet comme le R^2 ajusté de tenir compte de la parcimonie du modèle.

- **Press** : la statistique du Press (predicted residual error sum of squares) n'est affichée que si l'option correspondante a été activée dans la boîte de dialogue. Elle est définie par

$$Press = \sum_{i=1}^{n} w_i (y_i - \hat{y}_{i(-i)})^2$$

où $\hat{y}_{i(-i)}$ est la prédiction pour l'observation i lorsque cette dernière n'est pas utilisée pour l'estimation des paramètres. On obtient alors

$$\text{Press RMCE } = \sqrt{\frac{\text{Press}}{W - p^*}}$$

Le Press RMCE peut alors être comparé au RMCE. Une différence importante entre les deux indique que le modèle est sensible à la présence ou absence de certaines observations dans le modèle.

Le **tableau d'analyse de la variance** permet d'évaluer le pouvoir explicatif des variables explicatives. Dans le cas où la constante du modèle n'est pas fixée à une valeur donnée, le pouvoir explicatif est évalué en comparant l'ajustement (au sens des moindres carrés) du modèle final avec l'ajustement du modèle rudimentaire composé d'une constante égale à la moyenne de la variable dépendante. Dans le cas où la constante du modèle est fixée, la comparaison est faite par rapport au modèle pour lequel la variable dépendante serait égale à la constante fixée.

Dans le cas d'une régression PCR, le premier tableau des **paramètres du modèle** correspond aux paramètres du modèle s'appuyant sur les composantes principales sélectionnées. Ce tableau est difficilement interprétable. Pour cette raison, une transformation est opérée afin

d'obtenir les **paramètres du modèle** correspondant aux variables d'origine. Ce dernier tableau est obtenu directement dans le cas d'une régression OLS. Dans ce tableau, pour la constante du modèle et pour variable sont affichés l'estimation du paramètre, l'écart-type correspondant, le t de Student et la probabilité associée, ainsi que l'intervalle de confiance.

L'**équation du modèle** est ensuite affichée pour faciliter la lecture ou la réutilisation du modèle.

Le tableau des **coefficients normalisés** (aussi appelés coefficients bêta) permet de comparer le poids relatif des variables. Plus la valeur absolue d'un coefficient est élevée, plus le poids de la variable correspondante est important. Lorsque l'intervalle de confiance autour des coefficients normalisés comprend la valeur 0 (cela est facilement visible sur le graphique des coefficients normalisés), le poids d'une variable dans le modèle n'est pas significatif.

Dans le tableau des **prédictions et résidus** sont donnés pour chaque observation, son poids, la valeur de la variable explicative qualitative s'il n'y en a qu'une, la valeur observée de la variable dépendante, la prédiction du modèle, les résidus, les intervalles de confiance, ainsi que la prédiction ajustée. Deux types d'intervalles de confiance sont affichés : un intervalle de confiance autour de la moyenne (correspondant au cas où l'on ferait la prédiction pour un nombre infini d'observations avec un ensemble de valeurs données des variables explicatives) et un intervalle autour de la prédiction ponctuelle (correspondant au cas d'une prédiction isolée pour des valeurs données des variables explicatives). Le second intervalle est toujours plus grand que le premier, les aléas étant plus importants. Si des données de validation ont été sélectionnées, elles sont affichées en fin de tableau.

Les **graphiques** qui suivent permettent de visualiser les résultats mentionnés ci-dessus. S'il n'y a qu'une seule variable explicative dans le modèle, le premier graphique affiché permet de visualiser les valeurs observées, la droite de régression et les deux types d'intervalles de confiance autour des prévisions. Le second graphique permet quant à lui de visualiser les résidus normalisés en fonction de la variable explicative. En principe, les résidus doivent être distribués de manière aléatoire autour de l'axe des abscisses. L'observation d'une tendance ou d'une forme révèlerait un problème au niveau du modèle.

Les **trois graphiques** affichés ensuite permettent de visualiser respectivement l'évolution des résidus normalisés en fonction de la variable dépendante, la distance entre les prédictions et les observations (pour un modèle idéal, les points seraient tous sur la bissectrice), et les résidus normalisés sous la forme d'un diagramme en bâtons. Ce dernier graphique permet de rapidement voir si un nombre anormal de données sort de l'intervalle]-2, 2[sachant que ce dernier, sous hypothèse de normalité, doit contenir environ 95% des données.

Si vous avez sélectionné des données à utiliser pour calculer des **prédictions sur de nouvelles observations**, le tableau correspondant est ensuite affiché.

Dans le tableau des **diagnostics d'influence** sont affichés pour chaque observation, son poids, le résidu, le résidu normalisé (division par la RMCE), le résidu studentisé, le résidu supprimé, le résidu supprimé studentisé, le leverage centré, la distance de Mahalanobis, le D de Cook, le CovRatio, le DFFits, le DFFits standardisé, les DFBeta (un par coefficient du modèle) et les DFBeta standardisés.

Trois graphiques sont ensuite affichés pour mettre en évidence les observations dont l'influence nécessite une analyse particulière.

Exemple

Un exemple d'utilisation de la régression PLS sont disponibles sur le Centre d'aide XLSTAT aux adresses suivantes :

http://www.xlstat.com/demo-plsf.htm

Bibliographie

Akaike H. (1973). Information Theory and the Extension of the Maximum Likelihood Principle. In: Second International Symposium on Information Theory. (Eds: V.N. Petrov and F. Csaki). Academiai Kiadó, Budapest. 267-281.

Amemiya T. (1980). Selection of regressors. *International Economic Review*, **21**, 331-354.

Bastien P., Esposito Vinzi V. and Tenenhaus M. (2005). PLS generalised regression. *Computational Statistics and Data Analysis*, **48** , 17-46.

Dempster A.P. (1969). Elements of Continuous Multivariate Analysis. Addison-Wesley, Reading.

Kullback S. and Leibler R. A. (1951). On information and sufficiency. *Annals of Mathematical Statistics*, **22**, 79-86.

Ooms K. (1996). Identification of potentially causal regressors in PLS models. Dissertation: International Study Program in Statistics. KUL.

Schwarz G. (1978). Estimating the dimension of a model. *Annals of Statistics,* **6**, 461-464.

Tenenhaus M. **(1998)**. La Régression PLS, Théorie et Pratique. Technip, Paris.

Tenenhaus M., Pagès J., Ambroisine L. and Guinot C. (2005). PLS methodology for studying relationships between hedonic judgements and product characteristics. *Food Quality and Preference*. **16**, 4, 315-325.

Wold, S., Martens H. and Wold H. (1983). The Multivariate Calibration Problem in Chemistry solved by the PLS Method. In: Ruhe A.and Kågström B. (eds.), Proceedings of the Conference on Matrix Pencils. Springer Verlag, Heidelberg. 286-293.

Wold S. (1995). PLS for Multivariate Linear Modelling. In: van de Waterbeemd H. (ed.), QSAR: Chemometric Methods in Molecular Design. Vol 2. Wiley-VCH, Weinheim, Germany. 195-218.

Régression LASSO

Utilisez cet outil pour réaliser une régression lorsque vous avez plus de variables que d'observations ou, plus universellement, lorsque le nombre de variables est important.

Dans cette section :

Description

LASSO est l'acronyme de *Least Absolute Shrinkage and Selection Operator*. La régression LASSO a été proposée par Robert Tibshirani en 1996. C'est une méthode d'estimation qui contraint ses coefficients à ne pas *exploser*, contrairement à la régression linéaire standard en grande dimension. Le contexte de grande dimension recoupe toutes les situations où l'on dispose d'un très grand nombre de variables par rapport au nombre d'individus.

La régression LASSO est une des méthodes qui vient pallier les manques (instabilité de l'estimation et manque de fiabilité de la prévision) de la régression linéaire dans un contexte de grande dimension. L'avantage principal de la régression LASSO réside dans sa capacité à effectuer une sélection de variables, ce qui peut s'avérer précieux en présence d'un grand nombre de variables.

Régression LASSO

En notant Y le vecteur de la variable dépendante quantitative et X comme la matrice des variables explicatives, l'estimateur Lasso $\hat{\beta}$ est solution du problème de minimisation sous contrainte suivant :

$$ argmin_{\beta \in \mathbb{R}^{p}} L(\beta)=\left|Y-X\beta\right|^2 $$ sous la contrainte $\sum_{j=1}^{p} |\beta_j| \leq t$ pour un certain $t > 0$ et où p représente le nombre de variables.

Le Lagrangien associé au problème d'optimisation s'écrit :

$$||Y - X\beta||^2 + 2\lambda(\sum_{j=1}^{p}|\beta_j| - t)$$

où 2λ est le multiplicateur de Lagrange lié à t par la contrainte $\sum_{j=1}^{p} |\beta_j| = t$

En revanche, il n'existe pas de formule explicite pour la solution $\hat{\beta}$ à λ donné. De ce fait, la résolution du problème d'optimisation se fait grâce à des algorithmes. Dans XLSTAT, cette

résolution s'effectue à partir de l'algorithme de descente par coordonnée.

L'algorithme de descente par coordonnée

En utilisant la forme du Lagrangien, on exprime, pour $\lambda > 0$ donné, l'estimateur Lasso $\hat{\beta}$ comme étant $argmin_{\beta \in \mathbb{R}^p}$ de :

$$L(\beta) = \frac{1}{2} \|Y - X\beta\|_2^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

La régression LASSO se fonde sur un paramètre fondamental : le paramètre de régularisation $\lambda > 0$. XLSTAT propose à ses utilisateurs de trouver ce paramètre λ optimal par validation croisée.

Dans l'algorithme de descente par coordonnée, l'optimisation de chacun des paramètres se fait séparément (en maintenant tous les autres fixes).

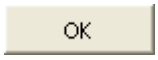
Ainsi, en notant que chaque variable X_j est centrée et réduite, la j-ème coordonnée $\hat{\beta}_j$ de la solution LASSO s'exprime de la façon suivante :

$$ \hat\beta_j = \left\lbrace \begin{aligned} & 0 & si & \lvert R_j \rvert \leq \lambda \ & (R_j-\lambda)/n & si & R_j > \lambda \ & (R_j+\lambda)/n & si & R_j < -\lambda \end{aligned} \right. $$ avec $R_j = X_j' Y - X_j' \sum_{k \neq j} X_k \hat{\beta}_k$ le j-ème résidu partiel, X_j' la transposée de la variable X_j et n le nombre d'observations.

L'algorithme de descente par coordonnée utilise cette formule pour mettre à jour chaque coordonnée de l'estimateur LASSO jusqu'à atteindre la convergence de L(β)

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les

variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.

 : cliquez sur ce bouton pour changer la façon dont vous chargez les données dans XLSTAT. Si le bouton possède une icône de souris, XLSTAT vous permet de faire une sélection à la souris des données. Si le bouton possède une icône de liste, XLSTAT vous permet de sélectionner les données à partir d'une liste. Si le bouton possède une icône orange, des boutons complémentaires avec un point d'interrogation  sont affichés pour vous permettre d'importer les données à partir de fichiers.

Onglet **Général** :

Y / Variables dépendantes :

Quantitatives : sélectionnez la ou les variables réponse que vous souhaitez modéliser. Si plusieurs variables sont sélectionnées, XLSTAT fera les calculs pour chacune des variables indépendamment. Si des en-têtes de colonnes ont été sélectionnés, veuillez vérifier que l'option « Libellés des variables » est activée.

Type de réponse : sélectionnez le type de réponse que vous avez :

- **Quantitative** : si votre variable réponse contient des valeurs numériques, choisissez ce type de variable réponse.

X / Variables explicatives :

Quantitatives : sélectionnez la ou les variables qualitatives explicatives sur la feuille Excel. Les données sélectionnées doivent être de type numérique. Si le libellé des variables a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des variables : activez cette option si la première ligne des données sélectionnées (variables, libellés des observations, poids des observations) contient un libellé.

Libellés des observations : activez cette option si vous voulez utiliser des libellés d'observations pour l'affichage des résultats. Si l'option « Libellés des variables » est activée, la première cellule de la sélection doit comprendre un en-tête. Si vous n'activez pas cette option, des libellés seront automatiquement créés (Obs1, Obs2, …).

Poids des observations : activez cette option si vous voulez pondérer les observations. Si vous n'activez pas cette option, les poids seront tous considérés comme valant 1. Les poids doivent être impérativement supérieurs ou égaux à 0. Un poids de 2 est équivalent à répéter deux fois la même observation. Si un en-tête de colonne a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

Onglet **Options** :

Validation croisée : activez cette option si vous souhaitez calculer le paramètre λ par validation croisée. Cette option vous permet de lancer une validation croisée "k-fold" pour obtenir le paramètre de régularisation λ optimal. Les données sont divisées en k blocs de taille égales. Un seul bloc est retenu en tant qu'échantillon de validation pour tester le modèle, et les k-1 blocs restants sont utilisés en tant qu'échantillon d'apprentissage.

- **Nombre de blocs** : entrez le nombre de blocs à constituer pour la validation croisée. Valeur par défaut : 5.
- **Nombre de valeurs testées** : entrez le nombre de valeurs de λ qui seront testées au cours de la validation croisée. Valeur par défaut : 100.

Lambda : activez cette option si vous souhaitez spécifier le paramètre de régularisation λ.

Conditions d'arrêt :

- **Convergence** : entrez la valeur seuil d'évolution maximale de log vraisemblance d'une itération à l'autre, qui une fois atteinte permet de considérer que l'algorithme a convergé. Valeur par défaut : 0,000001.
- **Temps maximum (en secondes)** : entrez le temps maximal alloué à une descente par coordonnée. Passé ce temps, si la convergence n'a pas été atteinte, l'algorithme s'arrête et renvoie les résultats obtenus lors de la dernière itération. Valeur par défaut : 180 secondes.

Interactions / Niveau : activez cette option pour inclure des interactions dans le modèle puis entrez le niveau maximum d'interaction (valeur comprise entre 1 et 5).

Onglet **Validation** :

Validation : activez cette option si vous souhaitez utiliser une partie des données sélectionnées pour valider le modèle.

Jeu de validation : choisissez l'une des options pour définir le mode de sélection des observations utilisées pour la validation :

- **Aléatoire** : les observations sont sélectionnées de manière aléatoire. Le « Nombre d'observations » doit alors être saisi.
- **N dernières lignes** : les N dernières observations sont sélectionnées pour la validation. Le « Nombre d'observations » N doit alors être saisi.
- **N premières lignes** : les N premières observations sont sélectionnées pour la validation. Le « Nombre d'observations » N doit alors être saisi.
- **Variable de groupe** : si vous choisissez cette option, vous devez ensuite sélectionner une variable indicatrice composée de 0 pour les observations à utiliser pour le calcul du modèle, et de 1 pour les observations à utiliser pour la validation du modèle.

Onglet **Prédiction** :

Prédiction : activez cette option si vous souhaitez sélectionner des données à utiliser en mode prédiction. Si vous activez cette option, vous devez veiller à ce que les données de prédiction soient organisées comme les données d'apprentissage : mêmes variables, même ordre dans les sélections.

Quantitatives : sélectionner la ou les variables quantitatives sur la feuille Excel. La première ligne doit comprendre l'en-tête si l'option libellés des variables est activée sur cet onglet.

Libellés des observations : activez cette option si vous voulez utiliser des libellés d'observations disponibles sur une feuille Excel pour l'affichage des résultats. Si vous n'activez pas cette option, des libellés seront automatiquement créés (PredObs1, PredObs2, …).

Libellés des variables : activez cette option si la première ligne des données sélectionnées (variables, libellés des observations) contient un libellé.

Onglet **Données manquantes** :

Ne pas accepter les valeurs manquantes : activez cette option pour que XLSTAT empêche la poursuite des calculs si des valeurs manquantes sont détectées.

Supprimer les observations : activez cette option pour supprimer les observations comportant des données manquantes.

Estimer les données manquantes : activez cette option pour estimer les données manquantes avant le début des calculs.

- **Moyenne ou mode** : activez cette option pour estimer les données manquantes en utilisant la moyenne (variables quantitatives) ou le mode (variables qualitatives) pour les variables correspondantes.
- **Plus proche voisin** : activez cette option pour estimer les données manquantes d'une observation en recherchant le plus proche voisin de l'observation.

Onglet **Sorties** :

Statistiques descriptives : activez cette option pour afficher les statistiques descriptives du jeu de données d'apprentissage pour les variables sélectionnées.

- **Echantillon de validation** : activez cette option pour afficher également les statistiques descriptives du jeu de données de validation pour les variables sélectionnées.
- **Echantillon de prédiction** : activez cette option pour afficher également les statistiques descriptives du jeu de données de prédiction pour les variables sélectionnées.

Matrice de corrélation : activez cette option pour afficher un aperçu des corrélations entre les différentes variables sélectionnées pour l'échantillon d'apprentissage.

Coefficients normalisés : activez cette option pour afficher les paramètres normalisés du modèle (coefficients bêta).

Prédictions et résidus : activez cette option pour afficher les prédictions et les résidus pour l'ensemble des observations.

Tous les coefficients : activez cette option pour afficher dans les résultats les variables associées à des coefficients nuls.

Onglet **Graphiques** :

Prédictions et résidus : activez cette option pour afficher les graphiques suivants :

- Variable dépendante versus résidus.
- Prédictions pour la variable dépendante versus résidus.
- Prédictions pour la variable dépendante versus variable dépendante.
- Graphique en bâtons des résidus.

Importance des variables : activez cette option pour afficher sous forme de graphique les mesures d'importance des variables.

Evolution de la MCE (Validation croisée) : activez cette option pour afficher sous forme de graphique l'évolution de la MCE en fonction du paramètre lambda.

Résultats

Statistiques descriptives : le tableau des statistiques descriptives présente des statistiques simples pour toutes les variables sélectionnées. Le nombre de valeurs manquantes, le nombre de valeurs non manquantes, la moyenne, l'écart-type sont affichés pour les variables quantitatives.

Matrice de corrélation : ce tableau est affiché afin de vous permettre d'avoir un aperçu des corrélations entre les différentes variables sélectionnées.

Coefficients d'ajustement : dans ce tableau sont affichées les statistiques relatives à l'ajustement du modèle de régression :

- **Observations** : le nombre d'observations prises en compte dans les calculs. Dans les formules présentées ci-dessous n désigne le nombre d'observations.
- **Somme des poids** : la somme des poids des observations prises en compte dans les calculs. Dans les formules présentées ci-dessous W désigne la somme des poids.
- **DDL** : le nombre de degrés de liberté pour le modèle retenu (ce nombre est égal au nombre de coefficients non-nuls dans le modèle).
- **R²** : le coefficient de détermination du modèle. La valeur de ce coefficient est comprise entre 0 et 1. XLSTAT le calcule comme suit :

$$R^2 = \frac{\sum_{i=1}^{n} w_i(\hat{y}_i - \bar{y}_i)^2}{\sum_{i=1}^{n} w_i(y_i - \bar{y}_i)^2} \text{ avec } \bar{y} = \frac{1}{W}\sum_{i=1}^{n} w_i y_i$$

Le R^2 s'interprète comme la proportion de la variabilité de la variable dépendante expliquée par le modèle. Plus le R^2 est proche de 1, meilleur est le modèle.

- **MCE** : la moyenne des carrés des erreurs (MCE) est définie par :

$$MCE = \frac{1}{W}\sum_{i=1}^{n} w_i(y_i - \hat{y}_i)^2$$

- **RMCE** : la racine de la moyenne des carrés des erreurs (RMCE) est la racine carrée de la MCE.

Paramètres du modèle : ce tableau fournit pour chaque paramètre sa valeur après ajustement du modèle

Coefficients normalisés : ce tableau des coefficients normalisés (aussi appelés coefficients bêta) permet, si la matrice contenant les variables explicatives n'a pas été centrée, de comparer le poids relatif des variables. Plus la valeur absolue d'un coefficient est élevée, plus le poids de la variable correspondante est important.

Graphique de l'importance des variables : la mesure d'importance calculée pour une variable donnée est la valeur absolue de son coefficient dans la régression.

Graphique de l'évolution de la MCE (Validation croisée) : ce graphique montre l'évolution de la MCE en fonction du paramètre lambda.

Prédictions et résidus : ce tableau fournit, pour chaque observation, la valeur observée de la variable dépendante, la prédiction du modèle et les résidus.

Graphiques des prédictions et résidus : ces graphiques permettent de visualiser les résultats mentionnés ci-dessus.

Exemple

Un tutoriel sur la façon d'utiliser la régression LASSO est disponible sur le Centre d'aide XLSTAT :

http://www.xlstat.com/demo-lassof.htm

Bibliographie

Frédéric Lavancier (2020). Statistique en grande dimension.

Jerome Friedman, Trevor Hastie et Rob Tibshirani (2008). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Volume 2.

Jerome Friedman, Trevor Hastie et Rob Tibshirani (2010). Regularization paths for generalized linear models via coordinate descent. In *Journal of Statistical Software* (Vol. 58).

Rob Tibshirani (1996). Regression Shrinkage and Selection via the LASSO. In *Journal of the Royal Society* (Vol. 58).

Régression Ridge

Utilisez cet outil pour réaliser une régression lorsque vous avez plus de variables que d'observations ou, plus universellement, lorsque le nombre de variables est important.

Dans cette section :

Description

La régression Ridge, méthode dérivée de la régularisation de Tikhonov, a été proposée par Hoerl et Kennard en 1970. C'est une méthode d'estimation qui contraint ses coefficients à ne pas *exploser*, contrairement à la régression linéaire standard en grande dimension. Le contexte de grande dimension recoupe toutes les situations où l'on dispose d'un très grand nombre de variables par rapport au nombre d'individus.

La régression Ridge est une des méthodes qui vient pallier les manques (instabilité de l'estimation et manque de fiabilité de la prévision) de la régression linéaire dans un contexte de grande dimension. La régression Ridge se démarque de la régression LASSO dans sa plus grande robustesse face aux jeux de données présentant une forte multicolinéarité.

Régression Ridge

En notant Y le vecteur de la variable dépendante quantitative et X comme la matrice des variables explicatives, l'estimateur Ridge $\hat{\beta}$ est solution du problème de minimisation sous contrainte suivant :

$$ argmin_{\beta \in \mathbb{R}^{p}} L(\beta)=\left|Y-X\beta\right|^2 $$ sous la contrainte $\sum_{j=1}^{p} (\beta_j)^2 \leq t$ pour un certain $t > 0$ et où p représente le nombre de variables.

Le Lagrangien associé au problème d'optimisation s'écrit :

$$\|Y - X\beta\|^2 + \lambda(\sum_{j=1}^{p}(\beta_j)^2 - t)$$

où λ est le multiplicateur de Lagrange lié à t par la contrainte $\sum_{j=1}^{p}(\beta_j)^2 = t$.

Contrairement à la régression LASSO, l'estimateur Ridge $\hat{\beta}$ a une forme explicite :

$$\hat{\beta} = (X'X + \lambda I_p)^{-1} X'Y$$

où I_p est la matrice identité d'ordre p.

Cependant, en grande dimension, l'inversion de la matrice $X'X + \lambda I_p$ peut s'avérer compliquée. De ce fait, la résolution du problème d'optimisation se fait grâce à des algorithmes. Dans XLSTAT, cette résolution s'effectue à partir de l'algorithme de descente par coordonnée.

L'algorithme de descente par coordonnée

En utilisant la forme du Lagrangien, on exprime, pour $\lambda > 0$ donné, l'estimateur Ridge $\hat{\beta}$ comme étant $argmin_{\beta \in \mathbb{R}^p}$ de :

$$L(\beta) = \|Y - X\beta\|_2^2 + \lambda \sum_{j=1}^{p} (\beta_j)^2$$

La régression Ridge se fonde sur un paramètre fondamental : le paramètre de régularisation $\lambda > 0$. XLSTAT propose à ses utilisateurs de trouver ce paramètre λ optimal par validation croisée.

Dans l'algorithme de descente par coordonnée, l'optimisation de chacun des paramètres se fait séparément (en maintenant tous les autres fixes).

Ainsi, en notant que chaque variable X_j est centrée et réduite, la j-ème coordonnée $\hat{\beta}_j$ de la solution Ridge s'exprime de la façon suivante :

$$ \hat\beta_j = \frac{R_j}{n(1+\lambda)} $$ avec $R_j = X_j'Y - X_j' \sum_{k \neq j} X_k \hat{\beta}_k$ le j-ème résidu partiel, X_j' la transposée de la variable X_j et n le nombre d'observations.

L'algorithme de descente par coordonnée utilise cette formule pour mettre à jour chaque coordonnée de l'estimateur Ridge jusqu'à atteindre la convergence de $L(\beta)$

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.

OK : cliquez sur ce bouton pour lancer les calculs.

Annuler : cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

Aide : cliquez sur ce bouton pour afficher l'aide.

 : cliquez sur ce bouton pour rétablir les options par défaut.

 : cliquez sur ce bouton pour effacer les sélections de données.

 : cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.

 : cliquez sur ce bouton pour changer la façon dont vous chargez les données dans XLSTAT. Si le bouton possède une icône de souris, XLSTAT vous permet de faire une sélection à la souris des données. Si le bouton possède une icône de liste, XLSTAT vous permet de sélectionner les données à partir d'une liste. Si le bouton possède une icône orange, des boutons complémentaires avec un point d'interrogation  sont affichés pour vous permettre d'importer les données à partir de fichiers.

Onglet **Général** :

Y / Variables dépendantes :

Quantitatives : sélectionnez la ou les variables réponse que vous souhaitez modéliser. Si plusieurs variables sont sélectionnées, XLSTAT fera les calculs pour chacune des variables indépendamment. Si des en-têtes de colonnes ont été sélectionnés, veuillez vérifier que l'option « Libellés des variables » est activée.

Type de réponse : sélectionnez le type de réponse que vous avez :

- **Quantitative** : si votre variable réponse contient des valeurs numériques, choisissez ce type de variable réponse.

X / Variables explicatives :

Quantitatives : activez cette option si vous voulez inclure une ou plusieurs variables explicatives quantitatives dans le modèle. Sélectionnez alors la ou les variables correspondantes sur la feuille Excel. Les données sélectionnées doivent être de type numérique. Si le libellé des variables a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

Qualitatives : activez cette option si vous voulez inclure une ou plusieurs variables explicatives qualitatives dans le modèle. Sélectionnez alors la ou les variables correspondantes sur la feuille Excel. Les données sélectionnées peuvent être de tout type, mais les données numériques sont automatiquement considérées comme nominales. Si le libellé des variables a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des variables : activez cette option si la première ligne des données sélectionnées (variables, libellés des observations, poids des observations) contient un libellé.

Libellés des observations : activez cette option si vous voulez utiliser des libellés d'observations pour l'affichage des résultats. Si l'option « Libellés des variables » est activée, la première cellule de la sélection doit comprendre un en-tête. Si vous n'activez pas cette option, des libellés seront automatiquement créés (Obs1, Obs2, …).

Poids des observations : activez cette option si vous voulez pondérer les observations. Si vous n'activez pas cette option, les poids seront tous considérés comme valant 1. Les poids doivent être impérativement supérieurs ou égaux à 0. Un poids de 2 est équivalent à répéter deux fois la même observation. Si un en-tête de colonne a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

Onglet **Options** :

Validation croisée : activez cette option si vous souhaitez calculer le paramètre λ par validation croisée. Cette option vous permet de lancer une validation croisée "k-fold" pour obtenir le paramètre de régularisation λ optimal. Les données sont divisées en k blocs de taille égales. Un seul bloc est retenu en tant qu'échantillon de validation pour tester le modèle, et les k-1 blocs restants sont utilisés en tant qu'échantillon d'apprentissage.

- **Nombre de blocs** : entrez le nombre de blocs à constituer pour la validation croisée. Valeur par défaut : 5.
- **Nombre de valeurs testées** : entrez le nombre de valeurs de λ qui seront testées au cours de la validation croisée. Valeur par défaut : 100.

Lambda : activez cette option si vous souhaitez spécifier le paramètre de régularisation λ.

Conditions d'arrêt :

- **Convergence** : entrez la valeur seuil d'évolution maximale de log vraisemblance d'une itération à l'autre, qui une fois atteinte permet de considérer que l'algorithme a convergé. Valeur par défaut : 0,000001.
- **Temps maximum (en secondes)** : entrez le temps maximal alloué à une descente par coordonnée. Passé ce temps, si la convergence n'a pas été atteinte, l'algorithme s'arrête et renvoie les résultats obtenus lors de la dernière itération. Valeur par défaut : 180 secondes.

Interactions / Niveau : activez cette option pour inclure des interactions dans le modèle puis entrez le niveau maximum d'interaction (valeur comprise entre 1 et 5).

Onglet **Validation** :

Validation : activez cette option si vous souhaitez utiliser une partie des données sélectionnées pour valider le modèle.

Jeu de validation : choisissez l'une des options pour définir le mode de sélection des observations utilisées pour la validation :

- **Aléatoire** : les observations sont sélectionnées de manière aléatoire. Le « Nombre d'observations » doit alors être saisi.

- **N dernières lignes** : les N dernières observations sont sélectionnées pour la validation. Le « Nombre d'observations » N doit alors être saisi.

- **N premières lignes** : les N premières observations sont sélectionnées pour la validation. Le « Nombre d'observations » N doit alors être saisi.

- **Variable de groupe** : si vous choisissez cette option, vous devez ensuite sélectionner une variable indicatrice composée de 0 pour les observations à utiliser pour le calcul du modèle, et de 1 pour les observations à utiliser pour la validation du modèle.

Onglet **Prédiction** :

Prédiction : activez cette option si vous souhaitez sélectionner des données à utiliser en mode prédiction. Si vous activez cette option, vous devez veiller à ce que les données de prédiction soient organisées comme les données d'estimation : mêmes variables, même ordre dans les sélections.

Quantitatives : activez cette option pour sélectionner la ou les variables quantitatives explicatives. La première ligne doit comprendre l'en-tête si l'option libellés des variables est activée sur cet onglet.

Qualitatives : activez cette option pour sélectionner la ou les variables qualitatives explicatives. La première ligne doit comprendre l'en-tête si l'option libellés des variables est activée sur cet onglet.

Libellés des observations : activez cette option si vous voulez utiliser des libellés d'observations disponibles sur une feuille Excel pour l'affichage des résultats. Si vous n'activez pas cette option, des libellés seront automatiquement créés (PredObs1, PredObs2, …).

Libellés des variables : activez cette option si la première ligne des données sélectionnées (variables, libellés des observations) contient un libellé.

Onglet **Données manquantes** :

Ne pas accepter les valeurs manquantes : activez cette option pour que XLSTAT empêche la poursuite des calculs si des valeurs manquantes sont détectées.

Supprimer les observations : activez cette option pour supprimer les observations comportant des données manquantes.

Estimer les données manquantes : activez cette option pour estimer les données manquantes avant le début des calculs.

- **Moyenne ou mode** : activez cette option pour estimer les données manquantes en utilisant la moyenne (variables quantitatives) ou le mode (variables qualitatives) pour les variables correspondantes.

- **Plus proche voisin** : activez cette option pour estimer les données manquantes d'une observation en recherchant le plus proche voisin de l'observation.

Onglet **Sorties** :

Statistiques descriptives : activez cette option pour afficher les statistiques descriptives du jeu de données d'apprentissage pour les variables sélectionnées.

- **Echantillon de validation** : activez cette option pour afficher également les statistiques descriptives du jeu de données de validation pour les variables sélectionnées.
- **Echantillon de prédiction** : activez cette option pour afficher également les statistiques descriptives du jeu de données de prédiction pour les variables sélectionnées.

Matrice de corrélation : activez cette option pour afficher un aperçu des corrélations entre les différentes variables sélectionnées pour l'échantillon d'apprentissage.

Coefficients normalisés : activez cette option pour afficher les paramètres normalisés du modèle (coefficients bêta).

Prédictions et résidus : activez cette option pour afficher les prédictions et les résidus pour l'ensemble des observations.

Onglet **Graphiques** :

Prédictions et résidus : activez cette option pour afficher les graphiques suivants :

- Variable dépendante versus résidus.
- Prédictions pour la variable dépendante versus résidus.
- Prédictions pour la variable dépendante versus variable dépendante.
- Graphique en bâtons des résidus.

Importance des variables : activez cette option pour afficher sous forme de graphique les mesures d'importance des variables.

Evolution de la MCE (Validation croisée) : activez cette option pour afficher sous forme de graphique l'évolution de la MCE en fonction du paramètre lambda.

Résultats

Statistiques descriptives : le tableau des statistiques descriptives présente des statistiques simples pour toutes les variables sélectionnées. Le nombre de valeurs manquantes, le nombre de valeurs non manquantes, la moyenne, l'écart-type sont affichés pour les variables quantitatives.

Matrice de corrélation : ce tableau est affiché afin de vous permettre d'avoir un aperçu des corrélations entre les différentes variables sélectionnées.

Coefficients d'ajustement : dans ce tableau sont affichées les statistiques relatives à l'ajustement du modèle de régression :

- **Observations** : le nombre d'observations prises en compte dans les calculs. Dans les formules présentées ci-dessous n désigne le nombre d'observations.

- **Somme des poids** : la somme des poids des observations prises en compte dans les calculs. Dans les formules présentées ci-dessous W désigne la somme des poids.

- **DDL** : le nombre de degrés de liberté pour le modèle retenu (ce nombre est égal au nombre de coefficients non-nuls dans le modèle).

- **R²** : le coefficient de détermination du modèle. La valeur de ce coefficient est comprise entre 0 et 1. XLSTAT le calcule comme suit :

$$R^2 = \frac{\sum_{i=1}^{n} w_i(\hat{y}_i - \bar{y}_i)^2}{\sum_{i=1}^{n} w_i(y_i - \bar{y}_i)^2} \text{ avec } \bar{y} = \frac{1}{W}\sum_{i=1}^{n} w_i y_i$$

Le R^2 s'interprète comme la proportion de la variabilité de la variable dépendante expliquée par le modèle. Plus le R^2 est proche de 1, meilleur est le modèle.

- **MCE** : la moyenne des carrés des erreurs (MCE) est définie par :

$$MCE = \frac{1}{W}\sum_{i=1}^{n} w_i(y_i - \hat{y}_i)^2$$

- **RMCE** : la racine de la moyenne des carrés des erreurs (RMCE) est la racine carrée de la MCE.

Paramètres du modèle : ce tableau fournit pour chaque paramètre sa valeur après ajustement du modèle

Coefficients normalisés : ce tableau des coefficients normalisés (aussi appelés coefficients bêta) permet, si la matrice contenant les variables explicatives n'a pas été centrée, de comparer le poids relatif des variables. Plus la valeur absolue d'un coefficient est élevée, plus le poids de la variable correspondante est important.

Graphique de l'importance des variables : la mesure d'importance calculée pour une variable donnée est la valeur absolue de son coefficient dans la régression.

Graphique de l'évolution de la MCE (Validation croisée) : ce graphique montre l'évolution de la MCE en fonction du paramètre lambda.

Prédictions et résidus : ce tableau fournit, pour chaque observation, la valeur observée de la variable dépendante, la prédiction du modèle et les résidus.

Graphiques des prédictions et résidus : ces graphiques permettent de visualiser les résultats mentionnés ci-dessus.

Exemple

Un tutoriel sur la façon d'utiliser la régression Ridge est disponible sur le Centre d'aide XLSTAT :

http://www.xlstat.com/demo-ridgef.htm

Bibliographie

Frédéric Lavancier (2020). Statistique en grande dimension.

Jerome Friedman, Trevor Hastie et Rob Tibshirani (2008). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Volume 2.

Jerome Friedman, Trevor Hastie et Rob Tibshirani (2010). Regularization paths for generalized linear models via coordinate descent. In *Journal of Statistical Software* (Vol. 33).

Rob Tibshirani (1996). Regression Shrinkage and Selection via the LASSO. In *Journal of the Royal Society* (Vol. 58).

Régression Elastic Net

Utilisez cet outil pour réaliser une régression lorsque vous avez plus de variables que d'observations ou, plus universellement, lorsque le nombre de variables est important.

Dans cette section :

Description

La régression Elastic net est un compromis entre les régressions Ridge et LASSO. C'est une méthode d'estimation qui contraint ses coefficients à ne pas *exploser*, contrairement à la régression linéaire standard en grande dimension. Le contexte de grande dimension recoupe toutes les situations où l'on dispose d'un très grand nombre de variables par rapport au nombre d'individus.

La régression Elastic net est une des méthodes qui vient pallier les manques (instabilité de l'estimation et manque de fiabilité de la prévision) de la régression linéaire dans un contexte de grande dimension. L'idée de cette méthode est de tirer partie des qualités de sélection de l'estimateur LASSO, tout en garantissant une meilleure robustesse en cas de multicolinéarité, propriété inhérente à la régression Ridge.

Régression Elastic net

En notant Y le vecteur de la variable dépendante quantitative et X comme la matrice des variables explicatives, l'estimateur Elastic net $\hat{\beta}$ est solution du problème de minimisation sous contrainte suivant :

$$ argmin_{\beta \in \mathbb{R}^{p}} L(\beta)=\left|Y-X\beta\right|^2 $$ sous la contrainte $\sum_{j=1}^{p}$ $ ((1-\alpha)\beta_j^2 + \alpha\lvert\beta_j \rvert)\leq t$ pour un certain $t > 0$ et où α est le paramètre de compromis (qui est compris entre 0 et 1) et p représente le nombre de variables.

Le Lagrangien associé au problème d'optimisation s'écrit :

$$\|Y - X\beta\|^2 + \lambda(\sum_{j=1}^{p}((1-\alpha)\beta_j^2 + \alpha|\beta_j|) - t)$$

où λ est le multiplicateur de Lagrange lié à t par la contrainte $\sum_{j=1}^{p}(1-\alpha)\beta_j^2 + \alpha|\beta_j| = t$

En revanche, il n'existe pas de formule explicite pour la solution $\hat{\beta}$ à λ donné. De ce fait, la résolution du problème d'optimisation se fait grâce à des algorithmes. Dans XLSTAT, cette résolution s'effectue à partir de l'algorithme de descente par coordonnée.

L'algorithme de descente par coordonnée

En utilisant la forme du Lagrangien, on exprime, pour $\lambda > 0$ donné, l'estimateur Elastic net $\hat{\beta}$ comme étant $argmin_{\beta \in \mathbb{R}^p}$ de :

$$L(\beta) = \|Y - X\beta\|_2^2 + \lambda \sum_{j=1}^{p}((1-\alpha)\beta_j^2 + \alpha|\beta_j|)$$

La régression Elastic net se fonde sur deux paramètres fondamentaux : le paramètre de compromis α (qui est compris entre 0 et 1) et le paramètre de régularisation $\lambda > 0$. XLSTAT propose à ses utilisateurs de trouver ces paramètres optimaux par validation croisée.

Dans l'algorithme de descente par coordonnée, l'optimisation de chacun des paramètres se fait séparément (en maintenant tous les autres fixes).

Ainsi, en notant que chaque variable X_j est centrée et réduite, la j-ème coordonnée $\hat{\beta}_j$ de la solution Elastic net s'exprime de la façon suivante :

$$\hat{\beta}_j = \begin{cases} 0 & si\ |R_j| \leq \lambda\alpha \\ (R_j - \lambda\alpha)/(n(1+\lambda(1-\alpha))) & si\ R_j > \lambda\alpha \\ (R_j + \lambda\alpha)/(n(1+\lambda(1-\alpha))) & si\ R_j < -\lambda\alpha \end{cases}$$

avec $R_j = X_j'Y - X_j' \sum_{k \neq j} X_k \hat{\beta}_k$ le j-ème résidu partiel, X_j' la transposée de la variable X_j et n le nombre d'observations.

L'algorithme de descente par coordonnée utilise cette formule pour mettre à jour chaque coordonnée de l'estimateur Elastic net jusqu'à atteindre la convergence de $L(\beta)$

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.

OK : cliquez sur ce bouton pour lancer les calculs.

Annuler : cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

Aide : cliquez sur ce bouton pour afficher l'aide.

 : cliquez sur ce bouton pour rétablir les options par défaut.

 : cliquez sur ce bouton pour effacer les sélections de données.

 : cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.

 : cliquez sur ce bouton pour changer la façon dont vous chargez les données dans XLSTAT. Si le bouton possède une icône de souris, XLSTAT vous permet de faire une sélection à la souris des données. Si le bouton possède une icône de liste, XLSTAT vous permet de sélectionner les données à partir d'une liste. Si le bouton possède une icône orange, des boutons complémentaires avec un point d'interrogation  sont affichés pour vous permettre d'importer les données à partir de fichiers.

Onglet **Général** :

Y / Variables dépendantes :

Quantitatives : sélectionnez la ou les variables réponse que vous souhaitez modéliser. Si plusieurs variables sont sélectionnées, XLSTAT fera les calculs pour chacune des variables indépendamment. Si des en-têtes de colonnes ont été sélectionnés, veuillez vérifier que l'option « Libellés des variables » est activée.

Type de réponse : sélectionnez le type de réponse que vous avez :

- **Quantitative** : si votre variable réponse contient des valeurs numériques, choisissez ce type de variable réponse.

X / Variables explicatives :

Quantitatives : sélectionnez la ou les variables qualitatives explicatives sur la feuille Excel. Les données sélectionnées doivent être de type numérique. Si le libellé des variables a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des variables : activez cette option si la première ligne des données sélectionnées (variables, libellés des observations, poids des observations) contient un libellé.

Libellés des observations : activez cette option si vous voulez utiliser des libellés d'observations pour l'affichage des résultats. Si l'option « Libellés des variables » est activée, la première cellule de la sélection doit comprendre un en-tête. Si vous n'activez pas cette option, des libellés seront automatiquement créés (Obs1, Obs2, …).

Poids des observations : activez cette option si vous voulez pondérer les observations. Si vous n'activez pas cette option, les poids seront tous considérés comme valant 1. Les poids doivent être impérativement supérieurs ou égaux à 0. Un poids de 2 est équivalent à répéter deux fois la même observation. Si un en-tête de colonne a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

Onglet **Options** :

Paramètres du modèle : cette option vous permet de décider de la méthode utilisée pour définir les paramètres du modèle. * **Validation croisée** : activez cette option si vous souhaitez calculer les paramètres du modèle par validation croisée. Cette option vous permet de lancer une validation croisée "k-fold" pour obtenir les paramètres optimaux du modèle. Les données sont divisées en k blocs de taille égales. Un seul bloc est retenu en tant qu'échantillon de validation pour tester le modèle, et les k-1 blocs restants sont utilisés en tant qu'échantillon d'apprentissage. * **Saisie manuelle** : activez cette option si vous souhaitez spécifier les paramètres du modèle.

Lambda : activez cette option si vous souhaitez calculer le paramètre λ par validation croisée. Dans le cas contraire, saisissez la valeur que vous souhaitez affecter au paramètre λ.

Alpha : activez cette option si vous souhaitez calculer le paramètre de compromis α par validation croisée. Dans le cas contraire, saisissez la valeur que vous souhaitez affecter au paramètre α.

Paramètres de la validation croisée : * **Nombre de blocs** : entrez le nombre de blocs à constituer pour la validation croisée. Valeur par défaut : 5. * **Nombre de valeurs testées** : entrez le nombre de valeurs de chacun des paramètres qui seront testées au cours de la validation croisée. Valeur par défaut : 100.

Conditions d'arrêt :

- **Convergence** : entrez la valeur seuil d'évolution maximale de log vraisemblance d'une itération à l'autre, qui une fois atteinte permet de considérer que l'algorithme a convergé. Valeur par défaut : 0,000001.

- **Temps maximum (en secondes)** : entrez le temps maximal alloué à une descente par coordonnée. Passé ce temps, si la convergence n'a pas été atteinte, l'algorithme s'arrête et renvoie les résultats obtenus lors de la dernière itération. Valeur par défaut : 180 secondes.

Interactions / Niveau : activez cette option pour inclure des interactions dans le modèle puis entrez le niveau maximum d'interaction (valeur comprise entre 1 et 5).

Onglet **Validation** :

Validation : activez cette option si vous souhaitez utiliser une partie des données sélectionnées pour valider le modèle.

Jeu de validation : choisissez l'une des options pour définir le mode de sélection des observations utilisées pour la validation :

- **Aléatoire** : les observations sont sélectionnées de manière aléatoire. Le « Nombre d'observations » doit alors être saisi.
- **N dernières lignes** : les N dernières observations sont sélectionnées pour la validation. Le « Nombre d'observations » N doit alors être saisi.
- **N premières lignes** : les N premières observations sont sélectionnées pour la validation. Le « Nombre d'observations » N doit alors être saisi.
- **Variable de groupe** : si vous choisissez cette option, vous devez ensuite sélectionner une variable indicatrice composée de 0 pour les observations à utiliser pour le calcul du modèle, et de 1 pour les observations à utiliser pour la validation du modèle.

Onglet **Prédiction** :

Prédiction : activez cette option si vous souhaitez sélectionner des données à utiliser en mode prédiction. Si vous activez cette option, vous devez veiller à ce que les données de prédiction soient organisées comme les données d'apprentissage : mêmes variables, même ordre dans les sélections.

Quantitatives : sélectionner la ou les variables quantitatives sur la feuille Excel. La première ligne doit comprendre l'en-tête si l'option libellés des variables est activée sur cet onglet.

Libellés des observations : activez cette option si vous voulez utiliser des libellés d'observations disponibles sur une feuille Excel pour l'affichage des résultats. Si vous n'activez pas cette option, des libellés seront automatiquement créés (PredObs1, PredObs2, …).

Libellés des variables : activez cette option si la première ligne des données sélectionnées (variables, libellés des observations) contient un libellé.

Onglet **Données manquantes** :

Ne pas accepter les valeurs manquantes : activez cette option pour que XLSTAT empêche la poursuite des calculs si des valeurs manquantes sont détectées.

Supprimer les observations : activez cette option pour supprimer les observations comportant des données manquantes.

Estimer les données manquantes : activez cette option pour estimer les données manquantes avant le début des calculs.

- **Moyenne ou mode** : activez cette option pour estimer les données manquantes en utilisant la moyenne (variables quantitatives) ou le mode (variables qualitatives) pour les variables correspondantes.
- **Plus proche voisin** : activez cette option pour estimer les données manquantes d'une observation en recherchant le plus proche voisin de l'observation.

Onglet **Sorties** :

Statistiques descriptives : activez cette option pour afficher les statistiques descriptives du jeu de données d'apprentissage pour les variables sélectionnées.

- **Echantillon de validation** : activez cette option pour afficher également les statistiques descriptives du jeu de données de validation pour les variables sélectionnées.
- **Echantillon de prédiction** : activez cette option pour afficher également les statistiques descriptives du jeu de données de prédiction pour les variables sélectionnées.

Matrice de corrélation : activez cette option pour afficher un aperçu des corrélations entre les différentes variables sélectionnées pour l'échantillon d'apprentissage.

Coefficients normalisés : activez cette option pour afficher les paramètres normalisés du modèle (coefficients bêta).

Prédictions et résidus : activez cette option pour afficher les prédictions et les résidus pour l'ensemble des observations.

Evolution de la MCE (Validation croisée) : activez cette option pour afficher sous forme de graphique l'évolution de la MCE en fonction des paramètres du modèle.

Tous les coefficients : activez cette option pour afficher dans les résultats les variables associées à des coefficients nuls.

Onglet **Graphiques** :

Prédictions et résidus : activez cette option pour afficher les graphiques suivants :

- Variable dépendante versus résidus.
- Prédictions pour la variable dépendante versus résidus.
- Prédictions pour la variable dépendante versus variable dépendante.
- Graphique en bâtons des résidus.

Importance des variables : activez cette option pour afficher sous forme de graphique les mesures d'importance des variables.

Evolution de la MCE (Validation croisée) : activez cette option pour afficher sous forme de graphique l'évolution de la MCE en fonction des paramètres du modèle.

Résultats

Statistiques descriptives : le tableau des statistiques descriptives présente des statistiques simples pour toutes les variables sélectionnées. Le nombre de valeurs manquantes, le nombre de valeurs non manquantes, la moyenne, l'écart-type sont affichés pour les variables quantitatives.

Matrice de corrélation : ce tableau est affiché afin de vous permettre d'avoir un aperçu des corrélations entre les différentes variables sélectionnées.

Coefficients d'ajustement : dans ce tableau sont affichées les statistiques relatives à l'ajustement du modèle de régression :

- **Observations** : le nombre d'observations prises en compte dans les calculs. Dans les formules présentées ci-dessous n désigne le nombre d'observations.
- **Somme des poids** : la somme des poids des observations prises en compte dans les calculs. Dans les formules présentées ci-dessous W désigne la somme des poids.
- **DDL** : le nombre de degrés de liberté pour le modèle retenu (ce nombre est égal au nombre de coefficients non-nuls dans le modèle).
- **R²** : le coefficient de détermination du modèle. La valeur de ce coefficient est comprise entre 0 et 1. XLSTAT le calcule comme suit :

$$R^2 = \frac{\sum_{i=1}^{n} w_i(\hat{y}_i - \bar{y}_i)^2}{\sum_{i=1}^{n} w_i(y_i - \bar{y}_i)^2} \text{ avec } \bar{y} = \frac{1}{W}\sum_{i=1}^{n} w_i y_i$$

Le R^2 s'interprète comme la proportion de la variabilité de la variable dépendante expliquée par le modèle. Plus le R^2 est proche de 1, meilleur est le modèle.

- **MCE** : la moyenne des carrés des erreurs (MCE) est définie par :

$$MCE = \frac{1}{W}\sum_{i=1}^{n} w_i(y_i - \hat{y}_i)^2$$

- **RMCE** : la racine de la moyenne des carrés des erreurs (RMCE) est la racine carrée de la MCE.

Paramètres du modèle : ce tableau fournit pour chaque paramètre sa valeur après ajustement du modèle

Coefficients normalisés : ce tableau des coefficients normalisés (aussi appelés coefficients bêta) permet, si la matrice contenant les variables explicatives n'a pas été centrée, de comparer le poids relatif des variables. Plus la valeur absolue d'un coefficient est élevée, plus le poids de la variable correspondante est important.

Graphique de l'importance des variables : la mesure d'importance calculée pour une variable donnée est la valeur absolue de son coefficient dans la régression.

Evolution de la MCE (Validation croisée) : ce tableau fournit l'évolution de la MSE en fonction des paramètres du modèle.

Graphique de l'évolution de la MCE (Validation croisée) : ce graphique montre l'évolution de la MCE en fonction des paramètres du modèle.

Prédictions et résidus : ce tableau fournit, pour chaque observation, la valeur observée de la variable dépendante, la prédiction du modèle et les résidus.

Graphiques des prédictions et résidus : ces graphiques permettent de visualiser les résultats mentionnés ci-dessus.

Exemple

Un tutoriel sur la façon d'utiliser la régression Elastic net est disponible sur le Centre d'aide XLSTAT :

https://www.xlstat.com/demo/elasticnet_fr

Bibliographie

Frédéric Lavancier (2020). Statistique en grande dimension.

Jerome Friedman, Trevor Hastie et Rob Tibshirani (2008). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Volume 2.

Jerome Friedman, Trevor Hastie et Rob Tibshirani (2010). Regularization paths for generalized linear models via coordinate descent. In *Journal of Statistical Software* (Vol. 58).

Rob Tibshirani (1996). Regression Shrinkage and Selection via the LASSO. In *Journal of the Royal Society* (Vol. 58).

Machine learning

Classification k-means floue

Utilisez la classification k-means floue pour constituer des groupes (classes) homogènes d'observations sur la base de leur description par un ensemble de variables quantitatives. Si le jeu de données possède plusieurs groupes d'observations très proches ou superposées, il est possible d'introduire un coefficient de flou (fuzziness) qui permet de relier chaque observation à un groupe avec une probabilité d'appartenance, c'est clustering flou ou soft clustering ou fuzzy clustering.

Dans cette section :

Description

Algorithme k-means

La classification k-means fut introduite par MacQueen en 1967. D'autres algorithmes similaires ont été développés par Forgy (1965) (centres mobiles). La classification k-means est une méthode itérative qui utilise des barycentres de classes lesquels sont mis à jour à chaque itération en associant à chaque barycentre son centre le plus proche. Chaque partition obtenue est caractérisée par une fonction objective Q à minimiser, l'algorithme continue d'effectuer des itérations jusqu'à ce que la différence entre deux fonctions objectives successives ($dQ = Q_{i-1} - Q_i$) soit inférieure à une valeur fixée par l'utilisateur. Le résultat de l'algorithme dépend du point de départ de celui-ci, donc en multipliant les points de départ et les répétitions on peut explorer différentes solutions possibles, ce qui réduit les chances de converger vers un optimum local. L'inconvénient de cette méthode est qu'elle ne permet pas de découvrir quel peut être un nombre cohérent de classes.

L'algorithme k-means, permet d'utiliser plusieurs indices de dissimilarité. La distance la plus utilisée est la distance euclidienne, mais dans certains cas comme l'analyse de texte, d'autres distances représentent mieux la structure du jeu de données comme la distance cosinus qui caractérise le k-means sphérique.

Le k-means sphérique est un dérivé de l'algorithme k-means qui permet de regrouper des observations selon l'angle les séparant (leurs tailles n'ont pas d'influence sur la distance). Cette classification est adaptée à l'analyse de texte où deux documents ayant la même proportion de mots mais en quantités différentes appartiennent à la même catégorie. Cette méthode à été mise au point par Dhillon en 2001 et permet en outre de faire un certain nombre d'optimisations algorithmiques en tirant partie des matrices creuses.

Les matrices de données issues d'analyses textuelles (matrices documents-termes) ne contiennent généralement que peu de valeurs positives. Ces dernières sont dites "creuses" en raison du nombre important de valeurs nulles qu'elles contiennent (au minimum 90% sur l'ensemble des observations). La structure de cette matrice peut être mise à profit en utilisant des conteneurs mémoires spécifiques permettant d'optimiser l'utilisation mémoire ainsi que la vitesse des calculs. XLSTAT utilise un conteneur spécifique pour matrice creuse de type matrice à lignes compressées qui ne conserve que les coordonnées et les valeurs non nulles de la matrice originale.

Algorithme k-means flou

La classification k-means floue permet de créer des classes d'observations dont les limites sont ambigües car trop proches les unes des autres. Cette méthode est apparue dès 1973 grâce aux travaux de Dunn et Bezdek et permet notamment de faire apparaître des sous-classes ou bien de faire une estimation du nombre de classes adéquat en faisant l'analyse sur un nombre de classes très élevé.

Le k-means flou est en fait une généralisation du k-means dans laquelle chaque observation possède une probabilité d'appartenir à chaque classe. On choisit lors de la première itération un point de départ qui consiste à associer le centre des k classes à k observations (prises au hasard ou non). On calcule ensuite la distance entre les observations et les k centres ainsi que la probabilité d'appartenance $\mu_{i,j}$ pour chaque observation i et chaque centre j :

$$\mu_{i,j} = \frac{\frac{1}{w_i d(X_i,C_j)}^{\frac{1}{m-1}}}{\sum_{l=1}^{k} \frac{1}{w_i d(X_i,C_l)}^{\frac{1}{m-1}}}$$

Puis on redéfinit les centres à partir des observations et de leurs probabilités d'appartenance modifiées par le coefficient de flou (ou de fuzziness) m.

$$C_j = \frac{\sum_{i=1}^{N} \sum_{j=1}^{k} w_i \mu_{i,j}^m X_i}{\sum_{i=1}^{N} \sum_{j=1}^{k} w_i \mu_{i,j}^m}$$

Le coefficient m doit être supérieur à 1, plus le coefficient m est grand et plus les frontières entres les classes sont floues (ATTENTION : Le coefficient est à choisir avec soin, un coefficient trop élevé risque de générer des probabilités d'appartenance égales et des classes n'ayant aucune signification). Le cas hard est un cas particulier du cas soft dans lequel $m = 1$ et si $j = min_j D(X_i, C_j)$ alors $\mu_{i,j} = 1$ sinon $\mu_{i,j} = 0$.

Indices de dissimilarité et critère de classification

Plusieurs indices de dissimilarité peuvent être utilisés pour parvenir à une solution. XLSTAT propose trois distances décritent par Chuanren Liu, Tianming Huy, Yong Gez et Hui Xiongx :

- **Distance Cosinus** : La distance cosinus est spécifique au k-means sphérique, elle est basée sur le cosinus de l'angle entre deux observations. Plus ce dernier est faible et plus la distance est petite, une distance de 1 correspond à deux observations n'ayant aucune valeur commune sur l'ensemble des variables.

$$D_{cosine}(A,B) = 1 - cos(A,B) = 1 - \frac{AB^T}{||A||||B||}$$

Dans le cas des analyses textuelles où la taille des différents documents ne doit avoir qu'une importance minime par rapport à leur contenu, ce type de distance est recommandé.

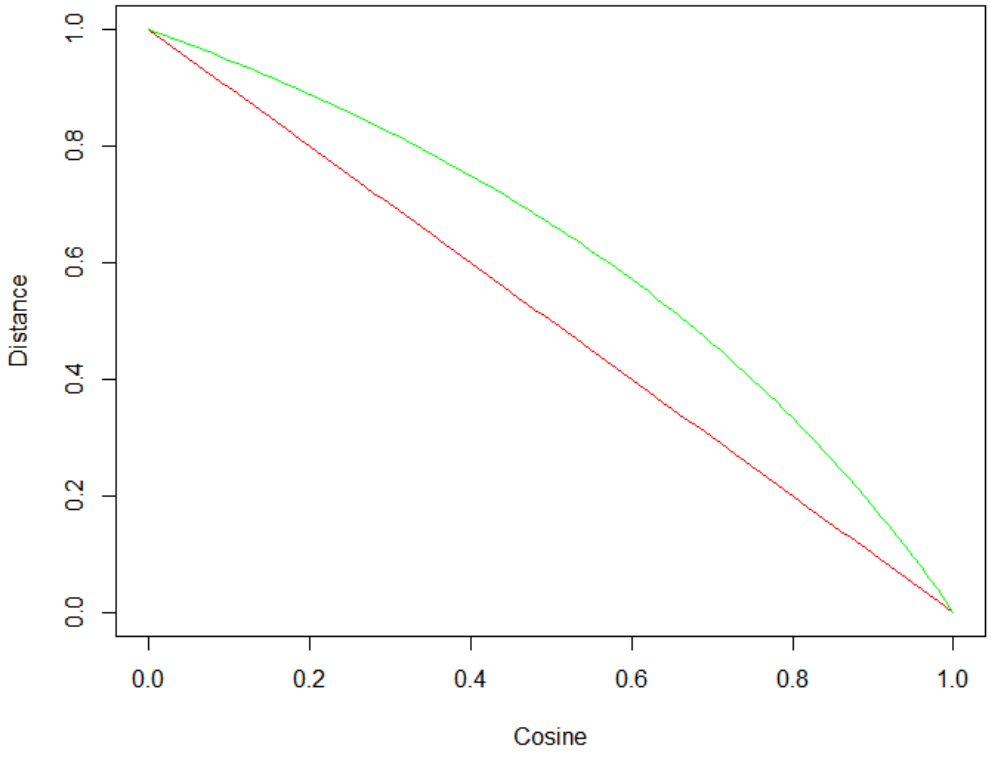
- **Distance de Jaccard** : Cette distance est basée sur l'indice de Jaccard étendu. Le coefficient de Jaccard compare l'ensemble des termes partagés à l'ensemble des termes qui sont présents dans chacun des deux documents mais ne sont pas les termes partagés.

$$D_{Jacc}(A,B) = 1 - Jacc(A,B) = 1 - \frac{A \cap B}{A \cup B}$$

L'index de Jaccard étendu effectue la même chose en traitant des observations non binaires. Pour des raisons d'optimisation, nous avons fondé le calcul de cette distance sur le cosinus.

$$D_{Jacc_{Extended}}(A,B) = 1 - Jacc_{Extended}(A,B) = 1 - \frac{cos(A,B)}{||A||^2 + ||B||^2 - cos(A,B)}$$

Par rapport à la distance cosinus, celle-ci possède un degré de sensibilité plus important à mesure que les deux observations sont proches, cela la rend utile dans des jeux de données denses contenant plusieurs classes rapprochées (Cf image ci-dessous).



Comparaison entre la distance cosinus (rouge) et la distance de Jaccard étendue (vert).

- **Distance Euclidienne** : La distance euclidienne est très commune en statistique et celle-ci permet d'obtenir des résultats robustes dans la plupart des cas. En revanche, pour des analyses impliquant des matrices creuses, il est recommandé d'utiliser les deux premières distances pour des raisons d'optimisation.

Le critère de classification Q (ou fonction objective) correspond à l'algorithme choisi : dans le cas euclidien trois choix sont possibles ($Trace(W)$, $Determinant(W)$, Lambda de Wilks), dans le cas de la distance de Jaccard c'est $Trace(W)$ et dans le cas sphérique c'est la somme des distances entre chaque observation et centre pondérés par μ et m(pour des raisons d'optimisation). Les critères sont présentés plus bas dans le document.

Q dans le cas sphérique :

$$Q = \sum_{i=1}^{N}\sum_{j=1}^{k}\mu_{i,j}^{m}D_{cosine}(X_i, C_j)$$

Dans le cas où la classification s'effectue avec une distance euclidienne, plusieurs critères de classification sont disponibles (Si la distance de Jaccard est utilisée c'est la $Trace(W)$ qui est le critère de classification) :

- **Trace(W)** : la trace de W, la matrice de variance intra-classe commune est le critère le plus classique. Minimiser la trace de W pour un nombre de classes donné, revient à minimiser la variance intra-classe totale, autrement dit à minimiser l'hétérogénéité des classes. Ce critère est sensible aux effets d'échelle. Si on ne veut pas donner plus de poids à certaines variables plutôt qu'à d'autres, on doit préalablement standardiser les données. Par ailleurs, ce critère tend à produire des classes de même taille.

Dans le cas hard :

$$W = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{k}\mu_{i,j}(X_i - C_j)(X_i - C_j)^T$$

Donc le critère est :

$$Q = \sum_{ii=1}^{D}W(ii, ii)$$

Dans le cas soft, la matrice W est décomposée en k matrice F_j :

$$F_j = \frac{1}{\sum_{i=1}^{N}\mu_{i,j}}\sum_{i=1}^{N}\mu_{i,j}(X_i - C_j)(X_i - C_j)^T$$

Ainsi, le critère devient

$$Q = \sum_{j=1}^{k}Trace(F_j) = \sum_{j=1}^{k}\sum_{ii=1}^{D}F_j(ii, ii)$$

- **Déterminant(W)** : le déterminant de W, la matrice de covariance intra-classe commune est un critère nettement moins sensible aux effets d'échelle que le critère $trace(W)$. Par ailleurs, la taille des classes peut être moins homogène qu'avec le critère de la trace.

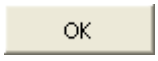
Dans le cas soft, le critère devient :

$$Q = \sum_{j=1}^{k} \sqrt{Determinant(F_j)}$$

- **Lambda de Wilks** : les résultats donnés par la minimisation de ce critère sont identiques à ceux donnés par la trace de W. Le critère du lambda de Wilks correspond à la division de $Trace(W)$ par $Trace(T)$ où T est la matrice de variance totale. La division par le déterminant de T permet d'avoir un critère toujours compris entre 0 et 1.

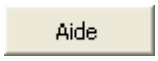
Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.



: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.

Onglet **Général** :

Tableau observations/variables : sélectionnez un tableau comprenant N observations décrits par P descripteurs. Si des en-têtes de colonnes ont été sélectionnés, veuillez vérifier que l'option « Libellés des variables » est activée.

Poids des observations : activez cette option si vous voulez pondérer les observations. Si vous n'activez pas cette option, les poids seront tous considérés comme valant 1. Les poids doivent être impérativement supérieurs ou égaux à 0. Si l'option "libellés des variables" est activée, la première cellule de la sélection doit comprendre un en-tête.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des variables : activez cette option si la première ligne des données sélectionnées (Tableau observations/variables, libellés des observations, poids des observations) contient un libellé.

Libellés des observations : activez cette option si vous voulez utiliser des libellés d'observations pour l'affichage des résultats. Si l'option « Libellés des variables » est activée, la première cellule de la sélection doit comprendre un en-tête. Si vous n'activez pas cette option, des libellés seront automatiquement créés (Obs1, Obs2, …).

Indice de dissimilarité : Choisissez parmi les trois distances proposées par XLSTAT. La distance cosinus est adaptée aux données textuelles. La distance de Jaccard est adaptée aux données qui requièrent une finesse d'analyse importante.

Critère de classification : Choisissez parmi les trois critères proposés par XLSTAT. La $trace(W)$ correspond à l'inertie intra-classe totale. Le Lambda de Wilks permet d'avoir un repère absolu du partitionnement. Le déterminant(W) est un critère nettement moins sensible aux effets d'échelle que le critère $trace(W)$.

Onglet **Options** :

Regrouper par observations : activez cette option si vous voulez créer des classes d'observations correspondant aux lignes et décrites par les colonnes correspondant aux variables.

Regrouper par variables : activez cette option si vous voulez créer des classes d'observations correspondant aux colonnes et décrites par les données correspondant aux lignes. C'est-à-dire, les variables sont traitées comme des observations dans ce cas-là.

Standardiser :

- **Centrer** : activez cette option si vous voulez centrer les données avant de commencer les calculs. Cette option soustrait la moyenne des variables aux données. La transformation sera appliquée aux variables.

- **Réduire** : activez cette option si vous voulez réduire les données avant de commencer les calculs. Cette option divise les données par l'écart-type des variables. Ceci permet de réduire les effets d'échelle dus à certaines variables dans la classification des données. La transformation sera appliquée aux variables.

Type de clustering :

- **Absolu** : Choisissez cette option pour effectuer l'algorithme classique du k-means (hard clustering).

- **Flou** : Choisissez cette option pour effectuer l'algorithme fuzzy k-means Le coefficient de flou par défaut est $1,05$.

Partition de départ : utilisez ces options pour choisir la manière dont est déterminée la partition initiale, autrement dit, la façon dont sont affectées les observations aux classes pour la première itération de l'algorithme de classification.

- **Aléatoire** : les observations sont affectées aux classes de manière aléatoire.

- **Définie par les centres** : l'utilisateur doit sélectionner les k centres correspondant aux k classes. Le nombre de lignes doit être égal au nombre de classes et le nombre de

colonnes au nombre de colonnes du tableau de données. Si l'option « Libellés des observations » est activée, la première cellule de la sélection doit comprendre un en-tête.

- **Définie par les appartenances** : les observations sont affectées aux classes suivant une variable indicatrice définie par l'utilisateur (ex : 2, 3, 6, 2 ,2, 5). Ce dernier doit dans ce cas sélectionner une variable indicatrice en colonne contenant autant de lignes que d'observations (avec éventuellement un en-tête).

- **K++** : cette option définit les centres initiaux en fonction de l'algorithme k-means++ développé par Rafail Ostrovsky, Yuval Rabani, Leonard Schulman et Chaitanya Swamy en 2006. Le premier centre est choisi aléatoirement parmi les observations. Le suivant est choisi parmi les observations en fonction de la distance entre l'observation et le centre. Plus la distance entre l'observation et le centre est grande et plus l'observation a de chance d'être sélectionnée. Les $k-2$ centres restants sont choisis en suivant la même méthode. Cette méthode permet de démarrer à partir de centres sélectionnés de manière homogène dans le jeu de données, ce qui conduit généralement à un nombre d'itérations moins grand et une qualité de partitionnement meilleure. En revanche, sur des données de grande taille et complexes (contenant beaucoup de centres), cet algorithme peut prendre un temps d'exécution non négligeable et il est préférable d'utiliser l'algorithme K|| .

- **K||** : cette option définit les centres en fonction de l'algorithme K|| ou "Scalable k-means" développé par Bahman Bahmani, Benjamin Moseley, Andrea Vattani, Ravi Kumar et Sergei Vassilvitskii en 2012. C'est une version modifiée de K++ qui permet d'effectuer le choix des centres initiaux en parallèle. Comme avec K++, le premier centre est choisi aléatoirement parmi les observations, puis à la prochaine itération , $k/2$ observations sont choisies de manière indépendante en fonction de leur distance par rapport au centre. Après un nombre d'itérations déterminé en fonction de la taille du jeu de donnée, les X centres ainsi obtenus sont agrégés en k centres via l'algorithme K++. Cet algorithme possède l'avantage d'être très rapide pour deux raisons : La première étape consiste à ne prendre qu'une partie des observations du jeu de données ce qui facilite grandement le travail de K++ et le choix de manière indépendante des centres à chaque itération permet de paralléliser une grande partie de l'algorithme.

Nombre de classes : entrez le nombre de classes qui doit être créé par l'algorithme. Vous pouvez choisir de faire varier le nombre de classes entre deux valeurs, sauf si "Définie par les centres" ou "Définie par les appartenances" sont choisis.

Nombre de répétitions : entrez le nombre de fois que l'algorithme est exécuté. Comme certains paramètres sont aléatoires, tels que les centres initiaux aléatoires par exemple, lancer plusieurs fois l'algorithme aide à atteindre une solution globale. Avec cette option, seule la meilleure classification sera prise en compte lors de l'affichage des résultats.

Conditions d'arrêt :

- **Itérations** : entrez le nombre maximal d'itérations pour l'algorithme k-means. Les calculs sont interrompus dès que le nombre maximal d'itérations est dépassé. Valeur par défaut : 50. Si la valeur du nombre d'itération est 0, l'algorithme itère jusqu'à convergence.

- **Convergence** : entrez la valeur minimale d'évolution du critère choisi d'une itération à l'autre, une fois cette valeur atteinte, on considère que l'algorithme à convergé. Valeur par

défaut : 0,00001. Si la valeur de convergence est 0, l'algorithme itérera jusqu'à atteindre le nombre d'itération maximal.

Onglet **Données manquantes** :

Ne pas accepter les valeurs manquantes : activez cette option pour que XLSTAT empêche la poursuite des calculs si des valeurs manquantes sont détectées.

Supprimer les observations : activez cette option pour supprimer les observations comportant des données manquantes.

Remplacer les valeurs par 0 : activez cette option pour remplacer les valeurs manquantes par 0.

Onglet **Sorties** :

Cet onglet est séparé en deux parties : les résultats globaux de toutes les partitions, et les résultats de chaque partition.

Résultats globaux

- **Tableau de synthèse** : activez cette option pour afficher la synthèse de l'optimisation. Ceci inclut le nombre de classes ainsi que d'itérations effectué par l'algorithme, le critère de classification, les variances inter et intra-classe, la largeur de la silhouette (Cf. description ci-dessous) ainsi que le lambda de Wilks.
- **Statistiques descriptives** : activez cette option pour afficher les statistiques descriptives pour les variables sélectionnées.
- **Effectif des classes** : activez cette option pour afficher le nombre d'observations par classe pour chaque partition créée.

Résultats par classe

- **Centres** : activez cette option pour afficher les coordonnées des centres des classes.
- **Objets centraux** : activez cette option pour afficher les coordonnées de l'observation la plus proche du centre de chaque classe.
- **Résumé du cluster** : activez cette option pour afficher les caractéristiques de chaque classe (variance intra-classe, distance moyenne, minimum et maximum) ainsi que les observations associées à chaque classe.
- **Colonnes les plus présentes** : activez cette option pour afficher les mots les plus présents dans chaque classe. Le nombre de mots affiché par défaut est 10.
- **Appartenances** : activez cette option pour afficher la classe à laquelle appartient chaque observation ainsi que la distance séparant l'observation du centre associé.
- **Probabilités d'appartenance** : Activez cette option pour afficher les probabilités d'appartenance $\mu_{i,j}$ associées à chaque observation de chacune des classes (en classification k-means floue uniquement)

Onglet **Graphiques** :

Évolution du critère : si vous avez choisi un nombre de classes entre deux bornes distinctes, XLSTAT affiche dans un premier temps l'évolution du critère de classification. Ce critère diminue lorsque le nombre de classes augmente. Si les données sont distribuées de manière homogène, la décroissance est linéaire. En revanche, si une structure de classe est bien présente, une zone de coude sera observée sur la courbe afin de déterminer le nombre adéquat de classes.

Profil des classes : activez cette option pour afficher un graphique permettant de comparer les moyennes des différentes classes créées.

Effectif des classes : activez cette option pour avoir une visualisation graphique de l'effectif des classes.

Silhouette : activez cette option pour afficher la silhouette de la partition. Pour chaque observation, on calcule un coefficient de fidélité allant de 1 à -1, où 1 correspond à une fidélité parfaite et un coefficient négatif correspond à une mauvaise partition (l'observation peut-être soit une valeur extrême ou bien alors la partition possède un mauvais nombre initial de classes). La formule du coefficient de fidélité est :

$$Fit(i) = \frac{b(i) - a(i)}{max(a(i), b(i)}$$

Où $a(i)$ est la distance moyenne entre l'observation i et les autres observations dans la même classe et $b(i)$ est le minimum des distances moyennes entre l'observation i et les autres observations d'une classe différente.

Silhouette condensé : activez cette option pour n'afficher qu'un nombre limité d'observations. Ceci augmente grandement la vitesse d'affichage et de navigation en cas de grand jeu de données. Cette option n'est prise en compte qu'à partir de 500 observations dans le jeu de données.

Exemple

Un exemple de classification k-means floue est disponible sur le Centre d'aide XLSTAT à l'adresse suivante :

http://www.xlstat.com/demo-fuzzyFR.htm

Bibliographie

MacQueen J. B. (1967). Some Methods for classification and Analysis of Multivariate Observations, *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, **Volume 1: Statistics**, 281-297

E.W. Forgy (1965). "Cluster analysis of multivariate data: efficiency versus interpretability of classifications, *Biometrics 21*, **3**, 768-769

S. Dhillon, Inderjit & S. Modha, Dharmendra. (2000). Concept Decompositions for Large Sparse Text Data Using Clustering, *Machine Learning*, **42**, 143-175.

C. Bezdek, James. (1981). Pattern Recognition with Fuzzy Objective Function.

Chuanren Liu, Tianming Huy, Yong Gez and Hui Xiongx. (2012). Which Distance Metric is Right: An Evolutionary K-Means View, SDM

Bahmani, Bahman & Moseley, Benjamin & Vattani, Andrea & Kumar, Ravi & Vassilvitskii, Sergei. (2012). Scalable K-Means++, *Proc. VLDB Endow*, **5(2012)**, 622–633

Rousseeuw, Peter. (1987). Rousseeuw, P.J.. Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis, *Journal of Computational and Applied Mathematics*, **20**, 53-65

K plus proches voisins

Utilisez cet outil pour prédire la valeur d'une observation pour une variable Y en fonction de la valeur pour cette même variable des k plus proches voisins, la proximité étant estimée sur un ensemble de variables X.

Dans cette section :

Description

La méthode des K plus proches voisins (KNN) a pour but de classifier des points cibles (classe méconnue) en fonction de leur distance par rapport à des points constituant un échantillon d'apprentissage (c'est-à-dire dont la classe est connue à priori).

KNN est une approche de classification supervisée intuitive. Il s'agit d'une généralisation de la méthode du plus proche voisin (NN). NN est un cas particulier de KNN où $k = 1$.

L'approche de classification KNN se base sur l'hypothèse que chaque cas de l'échantillon d'apprentissage est un vecteur aléatoire issu de $\mathbb{R}^n$. Chaque point est décrit comme $x = \{a_1(x), a_2(x), \ldots, a_n(x)\}$ où $a_r(x)$ correspond à la valeur du r-ème attribut. $a_r(x)$ peut être soit une variable quantitative, soit une variable qualitative.

Afin de déterminer la classe d'un point cible, chacun des k points les plus proches de x_q procèdent à un vote. La classe de x_q correspond à la classe majoritaire.

La méthode KNN de base peut être résumée suivant l'algorithme suivant :

Etant donné un jeu de données L de N échantillons pré-classifiés :

$$L = \{(x_1, f(x_1)), ..(x_2, f(x_2)), ..(x_N, f(x_N)\},$$

où $f(x_i)$ est une fonction de valeur réelle qui dénote la classe de x_i, $f(x_i) \in V$ avec $V = \{v_1, v_2, \ldots, v_s\}$,

- x_q est un point cible ou un échantillon à classifier.
- $x_1, x_2, \ldots, x_k$ sont les points dont la classe est connue et situés à une certaine distance de x_q.

- Alors :

$$f(x_q) = \mathrm{argmax}_{v \in V} \left(\sum_{i=1}^{K} \delta(v_i, f(x_i)) \right),$$

où $\delta(a,b) = \begin{cases} 1 \text{ si } a = b \\ 0 \text{ si } a \neq b \end{cases}$.

Origines

La méthode des plus proches voisins a d'abord été utilisée pour les reconnaissances de formes dans le cadre de sondages (Nilsson, 1965).

Elle a également été utilisée dans d'autres domaines, notamment :

- La bioinformatique
- Le traitement d'image
- Reconnaissance de formes (écriture manuscrite par exemple)
- Systèmes d'information géographique : trouver les villes les plus proches de certaines positions
- En général, dans les systèmes d'apprentissage, lorsque le problème implique la recherche des k points les plus proches d'un point cible donné.

Quantification de la similarité / dissimilarité entre un point cible et les points de l'échantillon d'apprentissage

La mesure de dissimilarité entre un point donné et les points de l'échantillon d'apprentissage se calcule grâce à une fonction de distance. Une fonction de distance d sur un échantillon X. d : $X \to R$ doit satisfaire les conditions métriques :

- $d(x,y) = d(y,x)$. Propriété de symétrie.
- $d(x,y) \geq 0$. Propriété de non-négativité.
- $d(x,y) = 0 \Leftrightarrow x = y$. Axiome de coïncidence.
- $d(x,y) = d(x,z) + d(z,y)$. Inégalité triangulaire.

Résultat asymptotique sur la convergence de l'algorithme KNN de base

Le résultat établi par Cover et Hart (1966) garantit l'existence des k plus proches voisins.

Soient x et $x_1, \ldots, x_N$ des variables aléatoires indépendantes identiquement distribuées pouvant prendre des valeurs au sein d'un espace métrique séparable X. Soit x'_n le voisin le plus proche de x au sein de l'ensemble $\{x_1, \ldots, x_N\}$.

Alors $x'_n \to x$ avec une probabilité de 1 (Cover et Hart 1966).

Complexité de la méthode KNN de base

Afin de trouver les K plus proches voisins d'un point donné, l'algorithme a besoin de calculer toutes les distances séparant le point-cible à chaque point dans l'échantillon d'apprentissage. Ainsi, l'algorithme calcule N distances, où N est le nombre de points au sein de l'échantillon d'apprentissage. Trouver les K plus proches voisins nécessite un tri de ces N distances. Ce tri constitue le goulot d'étranglement de l'algorithme. Par conséquent, la complexité de l'algorithme basique KNN est de l'ordre de $Nlog(N)$.

Métriques (distances) quantitatives :

Chaque point est considéré comme un vecteur quantitatif dont les composantes sont des variables aléatoires quantitatives.

Différentes distances quantitatives peuvent être utilisées :

- **Euclidienne** : $d(x,y) = \sum_{i=1}^{n} (x_i - y_i)^2$.
- **Minkowski** : $d(x,y) = \sum_{i=1}^{n} |x_i - y_i|^q$.
- **Manhattan** : $d(x,y) = \sum_{i=1}^{n} |x_i - y_i|$.
- **Tchebychev** : $d(x,y) = \max_{i=1.n}(|x_i - y_i|)$.
- **Canberra** : $d(x,y) = \sum_{i=1}^{n} \frac{|x_i - y_i|}{|x_i| + |y_i|}$.

Distances qualitatives :

Chaque point est perçu comme un vecteur dont les composantes sont des variables qualitatives. Dans ce cas, les distances quantitatives ne peuvent pas être utilisées. Différentes distances qualitatives ont été introduites :

La Distance d'Intersection

La distance d'intersection est une distance qualitative basique :

Deux vecteur x et y sont proches si leurs attributs sont similaires (s'ils portent les mêmes modalités/catégories).

La distance entre deux vecteurs peut être définie par :

$$d(x,y) = \sum_{i=1}^{N} \delta(a_i(x), a_i(y)),$$

où $a_i(x)$ et $a_i(y)$ correspondent aux i-ème attributs des vecteurs x et y.

La métrique de différences de valeurs (VDM)

VDM a été introduite par Stanfil et Waltz (1986). Deux attributs sont proches s'ils sont groupés au sein de la même classe. La distance **VDM** entre les vecteurs x et y est donnée par :

$$N1 : vdm_{normalisé_a}(x,y) = \sum_{c=1}^{C} |P(c|a_i(x)) - P(c|a_i(y))|^q$$

Où : - C est le nombre de classes de la variable réponse.

- $P(c|a_i(x))$: Sachant $a_i(x)$, la probabilité que $a_i(x)$ soit classé en c.
- $P(c|a_i(y))$: Sachant $a_i(y)$, la probabilité que $a_i(y)$ soit classé en c.
- q est généralement égal à 1 ou à 2.

$P(c|a_i(x))$ et $P(c|a_i(y))$ sont calculés de la sorte :

$$P(c|a_i(x)) = \frac{N(a_i,x,c)}{N(a_i,x)},$$
$$P(c|a_i(y)) = \frac{N(a_i,y,c)}{N(a_i,y)},$$

Où :

- $N(a_i,x,c)$: nombre de cas x avec a_i au sein de la classe c.
- $N(a_i,x)$: nombre de cas x au sein du jeu de données.
- $N(a_i,y,c)$: nombre de cas y avec a_i au sein de la classe c.
- $N(a_i,y)$: nombre de cas y au sein du jeu de données.

Remarque: bien que définie pour des attributs nominaux, la distance VDM peut également être utilisée pour évaluer une distance entre attributs quantitatifs.

Calcul de la similarité par la méthode des noyaux

Les noyaux sont une généralisation des mesures de distance. Ils peuvent être représentés par un espace de Hilbert (Scholkopf 2001).

Noyau Gaussien :

$$k(x,y) = \exp(-\frac{\|x-y\|}{2\delta^2})\cdot$$

Noyau Laplacien :

$$k(x,y) = \exp(-\frac{\|x-y\|}{\delta})\cdot$$

Noyau logarithmique :

$$k(x,y) = -\log(\|x-y\|^d + 1)\cdot$$

Noyau puissance :

$$k(x,y) = -\|x-y\|^d\cdot$$

Où x, y sont deux vecteurs de $\mathbb{R}^n$; δ et d sont des scalaires dans $\mathbb{R}$.

Noyau sigmoïde :

$$k(x,y) = \tanh(\alpha x^T y + c)\cdot$$

$x^T y$ est un produit scalaire

Noyau linéaire :

$$k(x,y) = \alpha x^T y + c,$$

où $x^T y$ est un produit scalaire

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.

OK : cliquez sur ce bouton pour lancer les calculs.

Annuler : cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.

Onglet **Général** :

Echantillon d'apprentissage

Y / Variables qualitatives

Sélectionnez la ou les variables réponses que vous souhaitez modéliser. Ces variables doivent être qualitatives. Si plusieurs variables sont sélectionnées, XLSTAT fera les calculs pour chacune des variables indépendamment. Si des en-têtes de colonnes ont été sélectionnés, veuillez vérifier que l'option « Libellés des variables » est activée.

X / Variables explicatives (échantillon d'apprentissage)

Quantitatives : sélectionnez la ou les variables explicatives quantitatives sur la feuille Excel. Les données sélectionnées doivent être de type numérique. Si le libellé des variables a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

Qualitatives : sélectionnez la ou les variables explicatives qualitatives (les facteurs) sur la feuille Excel. Les données sélectionnées peuvent être de tout type, mais les données numériques sont automatiquement considérées comme nominales. Si le libellé des variables a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des variables : activez cette option si la première ligne des données sélectionnées (variables dépendantes et explicatives, libellés des observations, poids) contient un libellé.

Onglet **Options** :

- Onglet **Général** :

Les options qui suivent sont liées à la procédure de classification KNN.

Modèle : sélectionnez le type de calcul (**métrique** ou **noyau**) pour calculer la similarité entre les points de l'échantillon de prédiction et les points de l'échantillon d'apprentissage.

Distance : sélectionnez le type de distance parmi celles proposées : **Canberra**, **Euclidienne**, **Minkowski**, **Manhattan**, **Tchebychev** sont des distances quantitatives. La Distance d'interception et la différence de valeurs (VDM) sont des distances qualitatives. Cette option est active en choisissant le modèle métrique.

Noyau : sélectionnez le type de noyau parmi ceux proposés : **Noyau gaussien**, **noyau laplacien**, **noyau sphérique**, **noyau linéaire**, **noyau puissance**. Cette option est active en choisissant le modèle noyau. Dans ce cas, les calculs sont plus longs à cause de la projection des points dans un espace à dimensionnalité plus importante.

```
**Prise en charge des égalités** :

Le vote à la majorité peut impliquer des égalités pour certains points (ex-
```

aequo).

```
La gestion des ex-aequo peut se faire selon différentes méthodes :

**Choix aléatoire** : cette option choisit la classe correspondant à un point
```

tiré aléatoirement dans l'échantillon des points équidistants.

```
**Plus petit indice** : cette option sélectionne la classe correspondant au
```

premier point rencontré dans l'échantillon de points équidistants.

```
**Voix pondérée** : cette option permet d'utiliser l'inverse de la distance
```

ou le carré de l'inverse de la distance en tant que poids associé aux votes des plus proches voisins.

```
**Observations suivies** : cette option permet d'explorer dans le détail les
```

k plus proches voisins pour certaines observations. Vous pouvez choisir **toutes** les observations ou uniquement certaines en indiquant au celles à retenir (1) et celles à ne pas garder (0).

- Onglet **Voisins** :

  Nombre de voisins : sélectionnez le nombre de voisins à utiliser durant la procédure KNN.

  - **Définies par l'utilisateur** : permet à l'utilisateur de définir manuellement le nombre de voisins, dans le champ **Nombre**.

- **Validation Croisée** :
  cette option permet de quantifier la qualité du classificateur KNN. La technique utilisée est la validation croisée « k-fold ». Les données sont divisées en k blocs de taille égale. Parmi les k blocs, un seul bloc est retenu en tant qu'échantillon de validation pour tester le modèle, et le reste des données est utilisé en tant qu'échantillon d'apprentissage. Puis on répète l'opération en sélectionnant un autre échantillon de validation parmi les k-1 échantillons qui n'ont pas encore été utilisés pour la validation du modèle. L'opération se répète ainsi k fois pour qu'en fin de compte chaque sous-échantillon ait été utilisé exactement une fois comme ensemble de validation. La moyenne des k erreurs est enfin calculée pour estimer l'erreur de prédiction. k peut être paramétré dans le champ **nombre de blocs**.

- **Automatique** : cette option permet à l'utilisateur de faire une recherche du nombre optimal de voisins dans une plage de valeurs à l'aide de la validation croisée. L'utilisateur choisit les bornes inférieure (champ **borne Inf.**) et supérieure (champ **borne sup.**) de l'intervalle, ainsi que le nombre de blocs à utiliser pour la validation croisée (champ **nombre de blocs**). Un tableau récapitulatif des erreurs de validation croisée obtenues lors de la recherche et le graphique associé sont affichés en sortie.

Onglet **Prédiction** :

Echantillon de prédiction :

sélectionnez les données relatives à l'échantillon de prédiction (quantitatives / qualitatives). Le nombre de variables doit être égal au nombre de variables explicatives de l'échantillon d'apprentissage.

Variables quantitatives : sélectionnez la ou les variables explicatives quantitatives sur la feuille Excel. Les données sélectionnées doivent être de type numérique. Si le libellé des variables a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

Variables qualitatives : sélectionnez la ou les variables explicatives qualitatives (les facteurs) sur la feuille Excel. Les données sélectionnées peuvent être de tout type, mais les données numériques sont automatiquement considérées comme nominales. Si le libellé des variables a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

Libellés des observations : activez cette option si vous voulez utiliser des libellés d'observations pour l'affichage des résultats. Si l'option « Libellés des variables » est activée, la première cellule de la sélection doit comprendre un en-tête. Si vous n'activez pas cette option, des libellés seront automatiquement créés (PredObs1, PredObs2, …).

Libellés des variables : activez cette option si la première ligne des données sélectionnées (variables explicatives, libellés des observations) contient un libellé.

Onglet **Données manquantes** :

Supprimer les observations : activez cette option pour supprimer les observations comportant des données manquantes.

Ignorer les données manquantes : activez cette option pour ignorer les observations comportant des données manquantes.

Estimer les données manquantes : activez cette option pour estimer les données manquantes avant le début des calculs.

- **Moyenne ou mode** : activez cette option pour estimer les données manquantes en utilisant la moyenne (variables quantitatives) ou le mode (variables qualitatives) pour les variables correspondantes.
- **Plus proche voisin** : activez cette option pour estimer les données manquantes d'une observation en recherchant le plus proche voisin de l'observation.

Onglet **Sorties**:

Statistiques descriptives : activez cette option pour afficher les statistiques descriptives pour les variables sélectionnées.

Résultats par classe : activez cette option pour afficher un tableau de statistiques ainsi que les objets pour chaque classe.

Résultats par objet : activez cette option pour afficher un tableau donnant la classe associée à chaque objet (observation) dans l'ordre initial des objets.

Résultats

Statistiques descriptives : les tableaux de statistiques descriptives présentent pour toutes les variables sélectionnées des statistiques simples. Pour les variables explicatives quantitatives, sont affichés le nombre d'observations, le nombre de données manquantes, le nombre de données non manquantes, la moyenne, et l'écart-type (non biaisé). Pour les variables dépendantes et les variables explicatives qualitatives sont affichés le nom des différentes modalités ainsi que leur fréquence respective.

Exemple

Un exemple de K-NN est disponible sur le Centre d'aide XLSTAT à l'adresse suivante :

http://www.xlstat.com/demo-knnf.htm

Bibliographie

Batista G. and Silva D. F. (2009). How k-Nearest Neighbor Parameters Affect its Performance?. Simposio Argentino de Inteligencia Artificial (ASAI 2009), 95-106.

Cover T.M. and Hart P.E. (1967). Nearest Neighbor pattern classification. *IEEE Transactions on Information Theory*, 13 (1):21-27.

Hechenbichler K. Schliep K. (2004). Weighted k-Nearest-Neighbor Techniques and Ordinal Classification. Sonderforschungsbereich 386, Paper 399.

Nilsson N (1965). Learning Machines. McGraw-Hill, New York.

Scholkopf B. (2001). The kernel trick distances. Advances in neural information processing systems. Microsoft Research, Redmond.

Sebestyen G. (1967). Decision-Making Processes in Pattern Recognition. Macmillan.

Stanfil G. and Walttz D. (1986). Towards memory based reasoning. *Communications of the ACM - Special issue on parallelism*, 29(12), 1213-1228.

Wilson D. R. (1972). Asymptotic Properties of Nearest Neighbor, Rules Using Edited Data. *IEEE Trans. On Systems Man and Cybernetics*, **2** (3), 408-421.

Wilson D. R. and Martinez T. R. (1997). Improved Heterogeneous Distance Functions. *Journal of Artificial Intelligence Research*, **6**, 1-34.

Classifieur bayésien naïf

Utilisez cet outil pour prédire une variable qualitative Y en fonction de variables explicatives qualitatives ou quantitatives.

Dans cette section :

Description

La méthode de classification naïve bayésienne est un algorithme d'apprentissage supervisé qui permet de classifier un ensemble d'observations selon des règles déterminées par l'algorithme lui-même. Cet outil de classification doit dans un premier temps être entrainé sur un jeu de données d'apprentissage qui montre la classe attendue en fonction des entrées. Pendant la phase d'apprentissage, l'algorithme élabore ses règles de classification sur ce jeu de donnée, pour les appliquer dans un second temps à la classification d'un jeu de données de prédiction. Le classificateur bayésien naïf implique que les classes du jeu de données d'apprentissage soient connues et fournit, d'où le caractère supervisé de l'outil.

Historiquement, la classification naïve bayésienne fut utilisée pour la classification de documents et l'élaboration de filtres anti-spam. Aujourd'hui, c'est un algorithme renommé dont les applications peuvent être rencontrées dans de nombreux domaines. Parmi ces atouts les plus significatifs, on citera son apprentissage rapide qui ne nécessite pas un gros volume de données et son extrême rapidité d'exécution comparé à d'autres méthodes plus complexes. Finalement, malgré la forte hypothèse simplificatrice d'indépendance des variables (voir description ci-dessous), la classification naïve bayésienne obtient des résultats remarquables dans de nombreuses applications de la vie courante ce qui en fait un algorithme de choix parmi les outils du machine learning.

A la base de la classification naïve bayésienne se trouve le théorème de Bayes avec l'hypothèse simplificatrice, dite naïve, d'indépendance entre toutes les paires de variables. Soit une classe de la variable y et un jeu de variables indépendantes $x_1, \ldots, x_n$, le théorème de bayes indique que :

$$P(y|x_1, \ldots, x_n) = \frac{P(y)P(x_1, \ldots, x_n|y)}{P(x_1, \ldots, x_n)}.$$

En utilisant l'hypothèse naïve d'indépendance des variables, on peut dériver la relation suivante :

$$P(x_i|y, x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n) = P(x_i|y).$$

On obtient donc, pour tout i, l'expression suivante :

$$P(y|x_1, \ldots, x_n) = \frac{P(y) \prod_{i=1}^{n} P(x_i|y)}{P(x_1, \ldots, x_n)}.$$

Comme $P(x_1, \ldots, x_n)$ reste le même pour toutes les classes, ce terme est considéré comme une constante de normalisation et la règle de classification devient la suivante :

$$P(y|x_1, \ldots, x_n) \propto P(y) \prod_{i=1}^{n} P(x_i|y),$$

et

$$\hat{y} = \max_y \left(P(y) \prod_{i=1}^{n} P(x_i|y) \right).$$

On peut utiliser une estimation de Maximum à Posteriori (MAP) pour estimer $P(y)$ et $P(x_i|y)$, où $P(y)$ est la fréquence relative de la classe y dans le jeu d'apprentissage.

Plusieurs types de classification naïve bayésienne peuvent être considérés en fonction de l'hypothèse qui est faite sur leur distribution conditionnelle $P(x_i|y)$.

$P(x_i|y)$ peut être considérée comme issue d'une distribution normale dont l'expression est :

$$P(x_i|y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2} \right).$$

Elle peut également être considérée comme issue d'une distribution de Bernoulli ou encore tout autre distribution paramétrique disponible dans XLSTAT : log-normale, gamma, exponentielle, logistic, Poisson, binomiale ou uniforme. Dans tous ces cas, les paramètres des distributions sont estimés en utilisant la méthode des moments.

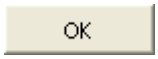
Si la distribution n'est pas connue ou bien si la variable considérée est de nature qualitative, XLSTAT propose d'estimer une distribution empirique à partir du rapport entre les observations correspondantes et le nombre total d'observation pour la classe y.

Si une distribution empirique est utilisée, un lissage de Laplace peut se reveler utile afin d'éviter une probabilité nulle. Ceci est particulièrement intéressant dans les cas où, par exemple, une variable qualitative du jeu de données de prédiction prend une valeur qui n'a pas été rencontrée lors de la phase d'apprentissage. La probabilité conditionnelle correspondante $P(x_i|y)$ prendrait alors la valeur de 0 pour toutes les classes y, aboutissant à un échec de la classification de l'observation. Les lissage de Laplace permet alors d'assigner une valeur très

faible mais différente de 0 à la probabilité conditionnelle $P(x_i|y)$, permettant ainsi aux autres variables d'être considérées pour réaliser la classification de l'observation.

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.

Onglet **Général** :

Echantillon d'apprentissage

Y / Variables qualitatives :

Sélectionnez la ou les variables réponses que vous souhaitez modéliser. Ces variables doivent être qualitatives. Si plusieurs variables sont sélectionnées, XLSTAT fera les calculs pour chacune des variables indépendamment. Si des en-têtes de colonnes ont été sélectionnés, veuillez vérifier que l'option « Libellés des variables » est activée.

X / Variables explicatives (échantillon d'apprentissage)

Quantitatives : sélectionnez la ou les variables explicatives quantitatives sur la feuille Excel. Les données sélectionnées doivent être de type numérique. Si le libellé des variables a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

Qualitatives : sélectionnez la ou les variables explicatives qualitatives (les facteurs) sur la feuille Excel. Les données sélectionnées peuvent être de tout type, mais les données numériques sont automatiquement considérées comme nominales. Si le libellé des variables a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des variables : activez cette option si la première ligne des données sélectionnées (variables dépendantes et explicatives, libellés des observations, poids) contient un libellé.

Onglet **Options** :

Distribution des variables quantitatives

Identique/Distribution paramétrique : Cette option vous permet de choisir la même distribution paramétrique parmi les distributions possibles pour toutes les variables quantitatives (voir ci-dessous la liste des distributions possibles).

Identique/Distribution empirique : Cette option vous permet de choisir la même distribution empirique pour toutes les variables quantitatives.

Spécifique : cette option permet de choisir une distribution spécifique à chaque variable quantitative.

Les distributions paramétriques peuvent être sélectionnées dans la liste suivante : normale, log-normale, gamma, exponentielle, logistic, Poisson, binomial, Bernoulli, uniforme.

Les variables qualitatives sont implicitement tirées à partir de distributions empiriques.

Les paramètres des distributions paramétriques sont estimés à partir de la méthode des moments.

Prise en charge des égalités :

La prédiction de la classification naïve bayésienne peut aboutir à un cas d'égalité où plusieurs classes obtiennent une même probabilité $P(y)$. Deux approches sont proposées pour gérer ces cas :

- **Choix aléatoire** : la classe est choisie de manière aléatoire dans l'ensemble des classes présentant la même probabilité $P(y)$.
- **Plus petit indice** : la classe choisie est la première classe rencontrée dans l'ensemble des classes présentant la même probabilité $P(y)$.

Paramètre de lissage :

Le lissage de Laplace permet d'éviter d'obtenir des probabilités nulles ou égales à un.

Le paramètre de lissage de Laplace θ est un entier positif ajouté lors du calcul de la fonction de probabilité $P(X_n = k)$ comme suit :

$$P(X_n = k) = \frac{n_k + \theta}{\sum_k n_k + \theta|V|},$$

avec X_n une variable qualitative ou quantitative. Le support de X_n est V supposé fini; la taille de l'ensemble V est $|V|$.

Onglet **Validation** :

La technique de validation utilisée pour vérifier la robustesse de la classification naïve est la validation croisée nommée **K-fold cross validation technique**. Les données sont divisées en **k** sous-ensembles de même taille. Parmi ces k sous-ensembles, un ensemble est mis de côté pour l'étape de validation et les $k-1$ sous-ensembles restants sont utilisés comme jeu d'apprentissage. La valeur de k peut être spécifiée dans le champ **Nombre de blocs**.

Onglet **Prédiction** :

Echantillon de prédiction :

Sélectionnez les données relatives à l'échantillon de prédiction (quantitatives / qualitatives). Le nombre de variables doit être égal au nombre de variables explicatives de l'échantillon d'apprentissage.

Variables quantitatives : sélectionnez la ou les variables explicatives quantitatives sur la feuille Excel. Les données sélectionnées doivent être de type numérique. Si le libellé des variables a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

Variables qualitatives : sélectionnez la ou les variables explicatives qualitatives (les facteurs) sur la feuille Excel. Les données sélectionnées peuvent être de tout type, mais les données numériques sont automatiquement considérées comme nominales. Si le libellé des variables a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

Libellés des observations : activez cette option si vous voulez utiliser des libellés d'observations pour l'affichage des résultats. Si l'option « Libellés des variables » est activée, la première cellule de la sélection doit comprendre un en-tête. Si vous n'activez pas cette option, des libellés seront automatiquement créés (PredObs1, PredObs2, …).

Libellés des variables : activez cette option si la première ligne des données sélectionnées (variables explicatives, libellés des observations) contient un libellé.

Onglet **Données manquantes** :

Supprimer les observations : activez cette option pour supprimer les observations comportant des données manquantes.

Estimer les données manquantes : activez cette option pour estimer les données manquantes avant le début des calculs.

- **Moyenne ou mode** : activez cette option pour estimer les données manquantes en utilisant la moyenne (variables quantitatives) ou le mode (variables qualitatives) pour les variables correspondantes.
- **Plus proche voisin** : activez cette option pour estimer les données manquantes d'une observation en recherchant le plus proche voisin de l'observation.

Onglet **Sorties** :

Statistiques descriptives : activez cette option pour afficher les statistiques descriptives pour les variables sélectionnées.

Résultats par classe : activez cette option pour afficher le tableau indiquant les statistiques et observations pour chaque classe.

Résultats par objet : activez cette option pour afficher le tableau indiquant la classe affectée à chaque observation dans l'ordre initial des observations.

Probabilités a posteriori : activez cette option pour afficher le tableau résumant les probabilités à posteriori correspondant à chaque classe $P(Y = y)$ pour toutes les observations prédites.

Matrice de confusion : activez cette option pour afficher la matrice de confusion. La matrice de confusion contient les informations concernant les classifications observées et prédites par l'algorithme. Les performances de l'algorithme peuvent être évaluées au moyen de cette matrice de confusion. La diagonale contient les prédictions correctes. Plus la somme des éléments de la diagonale est importante, meilleur est le classificateur.

La précision du modèle : activez cette option pour afficher la précision du modèle donnée par la proportion de prédictions correctes.

Résultats

Résultats correspondant aux statistiques descriptives de l'échantillon d'apprentissage

Le nombre d'observations correspondant à chaque variable dans l'échantillon d'apprentissage, sa moyenne (pour une variable quantitative, ses niveaux pour une variable qualitative) et sa déviation standard.

Résultats correspondant aux paramètres impliqués dans le processus de classification

Les distributions de probabilité utilisées sont indiquées.

Les variables qualitatives sont supposées suivre une distribution empirique.

La nature de la distribution *à priori* des classes (uniforme, non uniforme) est aussi rapportée.

Résultats concernant le classificateur

Afin d'évaluer et de noter le classificateur bayésien naïf, une matrice de confusion calculée avec la méthode du « leave one out » est indiquée.

Résultats concernant la méthode de validation

Le taux d'erreur de classificateur obtenu avec la validation croisée "K-folded" est indiquée. La valeur du paramètre K est également donnée.

Résultats concernant la prédiction des classes

Les classes prédites obtenues avec le classificateur bayésien naïf sont affichées. En plus des classes prédites, les probabilités à posteriori utilisées pour la prédiction sont également rapportées.

Exemple

Un exemple d'utilisation du Classifieur bayésien naïf est disponible sur le Centre d'aide XLSTAT à l'adresse suivante :

http://www.xlstat.com/demo-naivef.htm

Bibliographie

Abu Mustapha Y. S., MagDon-Ismail M., Lin H.-T. (2012). Learning From Data. AMLBook.

Mohri M., Rostamizadeh A., Talwalker A. (2012). Foundations of Machine Learning. MIT Press; Cambridge (Mass.).

Zhang H. (2004). The optimality of Naive Bayes. Proc. FLAIRS.

Machine à Vecteurs de Support

Utilisez cet outil pour réaliser une classification binaire, multi-classes ou une régression sur un échantillon d'observations décrites par des variables qualitatives et/ou quantitatives (prédicteurs).

Dans cette section :

Description

Les machines à vecteurs de support (SVM pour l'anglais Support Vector Machine) sont des outils d'apprentissage supervisé (Vapnik and Chervonenkis, 1964). Il fallut attendre le milieu des années 90 pour qu'une implémentation des SVM soit proposée avec l'introduction de l'astuce des noyaux (Boser, B., Guyon, I. & Vapnik V., 1992) et la généralisation au cas non séparable (Cortes C. & Vapnik V., 1995). Depuis lors, les SVM ont connu de nombreux développements et gagné en popularité dans divers domaines comme le machine learning, l'optimisation, les réseaux de neurones ou l'analyse fonctionnelle. C'est d'ailleurs l'un des algorithmes d'apprentissage qui a connu le plus de succès. Sa capacité à calculer des modèles complexes pour le coût calculatoire d'un modèle très simple en a fait une composante clef du domaine du machine learning où il s'est illustré en particulier dans la reconnaissance d'image ou de caractères.

Classification binaire

L'algorithme de SVM a pour objectif de trouver la séparation entre deux classes d'objets avec l'idée que plus la séparation est large, plus la classification est robuste. Dans sa forme la plus simple, celle d'une séparation linéaire et de classes séparables, l'algorithme sélectionne l'hyperplan qui sépare le jeu d'observations en deux classes distinctes de façon à maximiser la distance entre l'hyperplan et les observations les plus proches de l'échantillon d'apprentissage.

Supposons l'hyperplan P_0 optimal connu. Il est décrit par l'équation suivante :

$$P_0 : x^T \cdot w - b = 0$$

Avec x^T le jeu de prédicteurs de l'observation, w le vecteur normal à l'hyperplan et b l'origine de l'hyperplan.

L'échantillon d'apprentissage étant supposé séparable, nous pouvons identifier deux hyperplans, nommés P_+ et P_-, parallèles au plan de séparation de sorte que :

$$P_+ : x^T \cdot w - b = 1$$
$$P_- : x^T \cdot w - b = -1$$

Où $y_i = \pm 1$ indique les deux classes possibles de la sortie. La distance entre P_+ et P_-, $\frac{2}{\|w\|}$, est appelée marge. C'est le paramètre que nous souhaitons maximiser durant notre optimisation afin de nous assurer d'avoir la marge la plus large possible. Ce problème d'optimisation peut être formulé de la manière suivante :

$$\min_{w,b} \|w\|$$

sujet à :

$$y_i(x_i^T \cdot w - b) \geq 1, i = 1, ..., N$$

Les paramètres w et b sont donc obtenus à partir de la minimisation. Il est intéressant de noter que seuls les points proches de la limite influent sur l'hyperplan. Ces observations sont appelées vecteurs de support et revêtent un intérêt particulier car ils suffisent à définir notre classifieur. Au contraire, les observations éloignées de la limite ont seulement un effet marginal. Cette propriété peut se révéler intéressante dans des situations où des outliers sont présents.

Dans le cas où les classes de sorties se recouvrent, les données ne sont plus séparables linéairement et certains points doivent être autorisés à être du mauvais côté de la marge. Une variable ϵ_i est donc introduite pour rendre compte de la distance par laquelle la prédiction se trouve du mauvais côté.

Le problème d'optimisation devient alors :

$$min_{w,b} \|w\|$$

sujet à :

$$\begin{cases} y_i \left(x_i^T \cdot w - b\right) \geq 1 - \epsilon_i, \forall i \\ \epsilon_i \geq 0, \sum \epsilon_i \leq K \end{cases}$$

Les mauvaises classifications arrivent lorsque $\epsilon_i > 1$ et la somme totale des distances de mauvaise classification, $\sum \epsilon_i$, est bornée par une constante K.

En termes de calculs, il est plus aisé de reformuler l'expression ci-dessus dans la forme équivalente qui suit :

$$\min_{w,b} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{N} \epsilon_i$$

sujet à :

$$\begin{cases} y_1(x_i^T \cdot w - b) \geq 1 - \epsilon_i, \forall i \\ \epsilon_i \geq 0 \end{cases}$$

Où le paramètre de régularisation C est introduit en remplacement de K. Intuitivement, ce paramètre reflète le niveau de mauvaise classification autorisé. Une valeur large de C signifie que l'on souhaite limiter les mauvaises classifications au prix d'une marge plus étroite. Le cas extrême étant le cas séparable où $C = \infty$. Un C plus petit signifie que davantage de mauvaises classifications seront autorisées avec le bénéfice d'une marge plus large.

D'un point de vue calculatoire, le problème d'optimisation ci-dessus est quadratique avec des contraintes linéaires. Il peut donc être résolu en utilisant la méthode des multiplicateurs de Lagrange.

La fonction primale de Lagrange est :

$$L_p = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}\epsilon_i - \sum_{i=1}^{N} \propto_i \ [y_i(x_i^T w + b) - (1 - \epsilon_i)] - \sum_{i=1}^{N}\mu_i\epsilon_i$$

Que l'on va minimiser par rapport à w_i, b et ϵ_i.

Mettre les dérivées respectives à zéro donne :

$$\begin{aligned} w =& \sum_{i=1}^{N} \propto_i \ y_i x_i \\ 0 =& \propto_i \ y_i \\ \propto_i =& C - \mu_i \end{aligned}$$

Avec les contraintes positives $\propto_i, \mu_i, \epsilon_i \geq 0, \forall i$.

La fonction objective duale de Wolf à maximiser est :

$$L_D = \sum_{i=1}^{N} \propto_i \ -\frac{1}{2}\sum_{i=1}^{N}\sum_{i'=1}^{N} \propto_i \propto_{i'} \ y_i y_{i'} x_i^T x_{i'}$$

sujette à :

$$\begin{cases} 0 \leq \propto_i \leq C \\ \sum_{i=1}^{N} \propto_i \ y_i = 0 \end{cases}$$

Les conditions de Karush-Kuhn-Tucker (KKT) incluent les contraintes suivantes :

$$\begin{cases} \propto_i \ [y_i(x^T w + b) - (1 - \epsilon_i)] = 0 \\ \mu_i \epsilon_i = 0 \\ y_i(x^T w + b) - (1 - \epsilon_i) \geq 0 \end{cases}$$

pour $i = 1, ..., N$.

Dans le cas non séparable, les vecteurs de support sont identifiés comme étant les observations pour lesquelles $\propto_i > 0$. Pour celles situées exactement sur la bordure, on a $0 < \propto_i < c$, pour les autres $\propto_i = C$.

Le classifieur SVM peut maintenant être étendu à la classification non linéaire grâce à l'utilisation de noyaux. Cette méthode, connue sous le nom de l'astuce des noyaux, est similaire à celle utilisée en régression logistique dans laquelle la méthode linéaire est rendue plus flexible par l'élargissement de l'espace des données. Pour la méthode SVM, l'utilisation des noyaux permet d'agrandir très largement l'espace des données. Ainsi, des structures plus compliquées peuvent être détectées. XLSTAT propose 3 noyaux en plus de la méthode linéaire :

- Noyau puissance : $k(x_i, x_i') = (\gamma \cdot (x_i^T x_i') + coefficient)^{degré}$
- Noyau RBF : $k(x_i, x_i') = e^{-\gamma \|x_i - x_i'\|^2}$
- Noyau sigmoïde : $k(x_i, x_{i'}) = \tanh(\gamma \cdot (x_i^T x_i' + coefficient))$

Une fois que la fonction de base est choisie, la procédure est identique à celle décrite ci-dessus.

Le problème d'optimisation discuté ci-dessus est résolu en utilisant le Sequential Minimal Optimization (SMO) comme proposé par John Platt (Platt J.,1998). Cet algorithme sépare le problème principal en plusieurs sous-problèmes qui peuvent être résolus analytiquement. La charge de calcul s'en trouve fortement réduite ce qui fait du classifieur SVM, un classifieur très puissant avec un coût calculatoire très limité. Néanmoins une version du SMO avec une convergence plus rapide a été proposée par Fan *et al.* qui utilise l'information de Second Ordre (Fan, R., Chen, P. & Lin, C., 2005).

Classification multi-classes

Le classifieur SVM ne pouvant résoudre que des problèmes binaires, différentes approches ont été mises en place pour résoudre les problèmes multi-classes. Leur principe est toujours le même : transformer le problème multi-classes en plusieurs problèmes binaires. XLSTAT propose deux méthodes différentes.

La première stratégie proposée est celle du « un contre tous » (One versus All en anglais). Notons K le nombre de classes et ω_i la classe i avec $i \in \{1, \ldots, K\}$. La sortie f_i du classifieur i est entraînée en prenant la classe i comme classe positive contre toutes les autres classes qui deviennent ainsi désignées négatives. Pour une nouvelle observation x on assigne la classe i qui a retourné la plus grande valeur $f_i(x)$.

La seconde approche est appelée « un contre un » (One versus One with Max-wins voting en anglais). Un classifieur binaire est créé pour chaque paire de classes distinctes. En tout, $K(K-1)/2$ classifieurs sont créés. Notons C_{ij} le classifieur binaire prenant les observations de ω_i comme positives et celles de ω_j comme négatives. Pour une nouvelle observation x, si C_{ij} classe x dans ω_i alors on augmente le vote pour la classe ω_i de 1, sinon on augmente celui de la classe ω_j de 1. On réalise ce vote pour tous les classifieurs C_{ij} et on assigne à x la classe qui a obtenu le plus grand nombre de votes. Si une égalité apparaît, on procède à un tirage aléatoire.

Régression

La méthode SVM a été généralisée pour être appliquée à un problème de régression ou à la prédiction de séries temporelles. Soit $\{x_i, y_i\}$, l'ensemble d'apprentissage pour $i = 1, ..., N$ avec x le jeu de prédicteurs de l'observation et $y_i \in \mathbb{R}$.

Dans le cas linéaire le but est d'estimer f, avec au plus une déviation ϵ par rapport à la variable cible y. Elle est représentée par :

$$ f(x) = x ^ { T } .w + b $$ Avec w, le vecteur normal à l'hyperplan et b, l'origine de l'hyperplan.

Le problème d'optimisation peut être formulé ainsi :

$$min_{w,b} \frac{1}{2} \|w\|^2$$

sujette à :

$$\begin{cases} y_i - w^t \cdot x_i - b \leq \epsilon \\ w^t \cdot x_i + b - y_i \leq \epsilon \end{cases}$$

On introduit les variables de relâchement (slack variables en anglais) ξ et ξ^*, qui vont nous permettre d'autoriser les erreurs tout en gardant une platitude (flatness en anglais) pour la fonction f. On arrive à la formulation suivante :

$$min_{w,b} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{N} (\xi_i + \xi_i^*)$$

sujette à :

$$\begin{cases} y_i - w^t \cdot x_i - b \leq \epsilon + \xi_i \\ w^t \cdot x_i + b - y_i \leq \epsilon + \xi_i^* \\ \xi_i, \xi_i^* \geq 0 \end{cases}$$

Avec $C > 0$ le paramètre de régularisation. Plus C est grand et plus l'écart sera pénalisé. Si $C = 0$, il n'y a pas de pénalisation.

Après avoir introduit, les variables de relâchement et pour résoudre le problème précédent, on doit utiliser la fonction de perte ϵ-insensible, plus connu en anglais sous le nom de ϵ-insensitive loss function. Cette fonction $|\xi|_\epsilon$ est décrite par :

$$|\xi|_\epsilon = \begin{cases} 0 & si\ |\xi| \leq \epsilon \\ |\xi| - \epsilon & sinon \end{cases}$$

Comme pour la classification, on résout le problème d'optimisation ci-dessus en utilisant la méthode des multiplicateurs de Lagrange.

La fonction primale de Lagrange est :

$$L_p = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}(\xi_i + \xi_i^*) - \sum_{i=1}^{N} \propto_i [\epsilon + \xi_i - y_i + w^t \cdot x_i + b]$$
$$- \sum_{i=1}^{N} \propto_i^* [\epsilon + \xi_i^* + y_i + w^t \cdot x_i - b] - \sum_{i=1}^{N}(\eta_i\xi_i + \eta_i^*\xi_i^*)$$

Avec $\forall i,\ \propto_i, \propto_i^*, \eta_i, \eta_i^* \geq 0$ les multiplicateurs de Lagrange.

On minimise L par rapport à b, w_i, ξ_i et ξ_i^* et on met les dérivées respectives à zéro, qui nous donne :

$$\sum_{i=1}^{N}(\propto_i - \propto_i^*) = 0$$
$$w - \sum_{i=1}^{N}(\propto_i - \propto_i^*)x_i = 0$$
$$C - \propto_i - \eta_i = 0$$
$$C - \propto_i^* - \eta_i^* = 0$$

On peut éliminer les variables η et η^* en passant par :

$$\eta = C - \propto_i$$
$$\eta^* = C - \propto_i^*$$

Enfin, la fonction objective duale à maximiser est :

$$ \begin{aligned} L _{ D } &= - \frac { 1} { 2} \sum { i,j = 1} ^ { N }(\propto_{i} - \propto_{i}^{(*)} (\propto_{j}-\propto_{j})(x_{i}^{t} \cdot x_{j})) \

&-\epsilon\sum _{ i ^ = 1} ^ { N } (\propto { i }+ \propto _{ i} ^ { *}) + \sum { i ^ = 1} ^ { N }y _{ i } (\propto { i }+ \propto _ { i} ^ { *}) \end{aligned} $$

sujette à :

$$\begin{cases} \sum_{i=1}^{N}(\propto_i + \propto_i^*) = 0 \\ \propto_i, \propto_i^* \in [0, C] \end{cases}$$

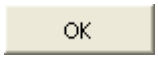
Les conditions de Karush-Kuhn-Tucker (KKT) sont les suivantes :

$$ \left\lbrace \begin{aligned} & \propto _{ i } [\epsilon + \xi+ w^{T} \cdot x_{i}+b] = 0 \ & \propto_ { i }^{} [\epsilon + \xi_{i}^{} + w^{t} \cdot x_{i}+b] = 0 \ & (C - \propto_{i}) \xi _ { i } = 0 \ & (C - \propto_{i}^{}) \xi _ { i }^{} = 0 \end{aligned} \right. $$ pour $i = 1, ..., N$.

Comme pour la classification, la méthode SVM pour la régression peut être étendue au cas non linéaire en utilisant l'astuce des noyaux. Quant à l'implémentation, elle se fait de même avec le Sequential Minimal Optimization (SMO) utilisant l'information de Second Ordre proposée par Fan *et al.*. (Fan, R., Chen, P. & Lin, C., 2005).

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.


: cliquez sur ces boutons pour changer la façon dont vous chargez les données dans XLSTAT. Si le bouton possède une icône de souris, XLSTAT vous permet de faire une sélection à la souris des données. Si le bouton possède une icône de liste, XLSTAT vous permet de sélectionner les données à partir d'une liste. Si le bouton possède une icône de feuille de papier orange, des nouveaux boutons s'affichent pour charger un fichier texte ou CSV en mémoire  .

Onglet **Général** :

Variable réponse : sélectionnez la variable réponse que vous souhaitez modéliser. Si des en-têtes de colonnes ont été sélectionnés, veuillez vérifier que l'option « Libellés des variables » est activée.

Type de réponse : sélectionnez le type de réponse que vous avez :

- **Qualitative** : sélectionnez ce type, si vous voulez entraîner un classifieur. Puis choisissez le type de classification : **binaire**, si vous avez sélectionné une variable contenant exactement deux valeurs distinctes. La classe positive correspond à la première catégorie rencontrée dans le jeu de données et la classe négative à la seconde. Si vous avez plus de deux classes, choisissez entre les méthodes **un contre un** et **un contre tous** qui correspondent à des méthodes multi-classes (voir la section Description).

- **Quantitative** : si votre variable réponse contient des valeurs numériques, choisissez ce type de variable réponse.

Variables explicatives :

Quantitatives : activez cette option si vous voulez inclure une ou plusieurs variables explicatives quantitatives dans le modèle. Sélectionnez alors la ou les variables correspondantes sur la feuille Excel. Les données sélectionnées doivent être de type numérique. Si le libellé des variables a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

Qualitatives : activez cette option si vous voulez inclure une ou plusieurs variables explicatives qualitatives dans le modèle. Sélectionnez alors la ou les variables correspondantes sur la feuille Excel. Les données sélectionnées peuvent être de tout type, mais les données numériques sont automatiquement considérées comme nominales. Si le libellé des variables a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des variables : activez cette option si la première ligne des données sélectionnées (variables dépendantes et explicatives, libellés des observations, poids) contient un libellé.

Libellés des observations : activez cette option si vous voulez utiliser des libellés d'observations pour l'affichage des résultats. Si l'option « Libellés des variables » est activée, la première cellule de la sélection doit comprendre un en-tête. Si vous n'activez pas cette option, des libellés seront automatiquement créés (Obs1, Obs2, …).

Poids des observations : activez cette option si vous voulez utiliser des poids différents pour des observations, en particulier lorsqu'une classe est majoritairement plus présente que la seconde. Si un en-tête de colonne a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

Onglet **Options** :

Paramètres SMO : cette option vous permet de régler l'algorithme d'optimisation selon vos besoins spécifiques. Il y a 3 paramètres ajustables :

- C : le paramètre de régularisation (voir la description pour plus de détails).
- **Tolérance** : ce paramètre définit la tolérance appliquée lors de la comparaison de deux valeurs durant la phase d'optimisation. Ce paramètre peut être utilisé pour accélérer la vitesse de calcul.
- **Epsilon** : utilisé seulement pour la régression, ce paramètre définit le tube d'insensibilité de rayon ϵ lié à la fonction de perte ϵ-insensible (voir la description pour plus de détails).

NB : le paramètre ϵ, présenté dans la classification binaire, est un paramètre de précision numérique dépendant de la machine est initialisé à 10^{-12}.

Prétraitement : cette option vous permet de sélectionner le prétraitement appliqué aux variables explicatives. Il y a 3 options disponibles :

- **Homothétie** : les variables explicatives quantitatives sont ajustées à une échelle allant de 0 à 1 en utilisant le minimum et maximum observé pour chaque variable.
- **Normalisation** : les variables explicatives quantitatives et qualitatives sont normalisées en utilisant la moyenne et la variance observée pour chaque variable.
- **Aucun** : aucun prétraitement n'est appliqué.

Validation croisée : cette option vous permet de lancer une validation croisée "k-fold" pour mesurer la qualité du classifieur ou de la régression avec les paramètres choisis. Les données sont divisées en k blocs de taille égale. Un seul bloc est retenu en tant qu'échantillon de validation pour tester le modèle, et les k-1 blocs restants sont utilisés en tant qu'échantillon d'apprentissage.

Noyau : cette option vous permet de sélectionner le noyau que vous souhaitez appliquer pour augmenter les dimensions de votre espace. Il y a 4 noyaux disponibles :

- **Noyau linéaire** : c'est le produit scalaire.
- **Noyau puissance** : ce noyau est détaillé dans la description. Si vous sélectionnez ce noyau, vous devez saisir la valeur du degré, du coefficient et de Gamma.
- **Noyau RBF** : c'est le noyau RBF présenté dans la description. Si vous sélectionnez ce noyau, vous devez saisir la valeur de Gamma.
- **Noyau sigmoïde** : ce noyau est détaillé dans la description. Si vous sélectionnez ce noyau vous devez saisir la valeur du coefficient et de Gamma.

Onglet **Validation** :

Validation : activez cette option si vous souhaitez utiliser une partie des données sélectionnées pour valider le modèle.

Jeu de validation : choisissez l'une des options pour définir le mode de sélection des observations utilisées pour la validation :

- **Aléatoire** : les observations sont sélectionnées de manière aléatoire. Le « Nombre d'observations » doit alors être saisi.
- **N dernières lignes** : les N dernières observations sont sélectionnées pour la validation. Le « Nombre d'observations » N doit alors être saisi.
- **N premières lignes** : les N premières observations sont sélectionnées pour la validation. Le « Nombre d'observations » N doit alors être saisi.

- **Variable de groupe** : si vous choisissez cette option, vous devez ensuite sélectionner une variable indicatrice composée de 0 pour les observations à utiliser pour le calcul du modèle, et de 1 pour les observations à utiliser pour la validation du modèle.

Onglet **Prédiction** :

Prédiction : activez cette option si vous souhaitez sélectionner des données à utiliser en mode prédiction. Si vous activez cette option, vous devez veiller à ce que les données de prédiction soient organisées comme les données d'estimation : mêmes variables, même ordre dans les sélections.

Quantitatives : activez cette option pour sélectionner la ou les variables quantitatives explicatives. La première ligne doit comprendre l'en-tête si l'option libellés des variables est activée sur cet onglet.

Qualitatives : activez cette option pour sélectionner la ou les variables qualitatives explicatives. La première ligne doit comprendre l'en-tête si l'option libellés des variables est activée sur cet onglet.

Libellés des observations : activez cette option si vous voulez utiliser des libellés d'observations disponibles sur une feuille Excel pour l'affichage des résultats. Si vous n'activez pas cette option, des libellés seront automatiquement créés (PredObs1, PredObs2, …).

Libellés des variables : activez cette option si la première ligne des données sélectionnées (variables explicatives, libellés des observations) contient un libellé.

Onglet **Données manquantes** :

Ne pas accepter les valeurs manquantes : activez cette option pour que XLSTAT empêche la poursuite des calculs si des valeurs manquantes sont détectées.

Supprimer les observations : activez cette option pour supprimer les observations comportant des données manquantes.

Estimer les données manquantes : activez cette option pour estimer les données manquantes avant le début des calculs.

- **Moyenne ou mode** : activez cette option pour estimer les données manquantes en utilisant la moyenne (variables quantitatives) ou le mode (variables qualitatives) pour les variables correspondantes.

- **Plus proche voisin** : activez cette option pour estimer les données manquantes d'une observation en recherchant le plus proche voisin de l'observation.

Onglet **Sorties** :

Statistiques descriptives : activez cette option pour afficher les statistiques descriptives pour les variables sélectionnées et pour chaque jeu de données présent (apprentissage, validation, prédiction).

Résumé de l'estimation : activez cette option pour afficher un résumé du classifieur ou de la régression SVM optimisé(e).

Liste des vecteurs de support : activez cette option pour afficher la liste complète des vecteurs de support et leurs coefficients associés α pour la classification ou α-α^* pour la régression, comme présentés dans la description.

Résultats par objet : activez cette option pour afficher les résultats de la classification ou de la régression pour chaque observation de l'échantillon d'apprentissage, de validation et de prédiction (si activé).

Indicateurs de performance : activez cette option pour afficher les indicateurs de performance pour la classification ou la régression de l'échantillon d'apprentissage et de validation (si un jeu de validation est activé).

Matrice de confusion : cette option est uniquement disponible pour la classification binaire ou multi-classes. Elle permet d'afficher la matrice de confusion des résultats de prédiction sur les échantillons d'apprentissage et de validation. La matrice de confusion contient les informations concernant les classifications observées et prédites par l'algorithme. Les performances de l'algorithme peuvent être évaluées au moyen de cette matrice de confusion. La diagonale contient les prédictions correctes. Plus la somme des éléments de la diagonale est importante, meilleur est le classifieur.

Résultats

Statistiques descriptives : le tableau des statistiques descriptives présente des statistiques simples pour toutes les variables sélectionnées. Le nombre de valeurs manquantes, le nombre de valeurs non manquantes, la moyenne, l'écart standard sont affichées pour les variables quantitatives. Pour les variables qualitatives, les catégories avec leur fréquence respective et pourcentage sont affichées.

Résultats associés à l'estimation :

Un résumé descriptif de l'estimation du classifieur optimisé ou de la régression est affiché. Dans le cas de la classification, les classes positive et négative sont indiquées. Dans les deux cas, la taille effective de l'échantillon d'apprentissage et les deux paramètres optimisés, le biais et le nombre de vecteurs de support sont affichés.

Résultats associés à la liste des vecteurs de support :

Un tableau, contenant la valeur optimisée de α ou α-α^* pour la régression et les variables explicatives prétraitées comme elles sont utilisées durant l'optimisation, est affiché. La taille du tableau dépend du nombre de vecteurs de support identifiés.

Résultats associés aux matrices de confusion :

Les matrices de confusion sont déduites des classifications obtenues et de la classe effective ainsi que les pourcentages d'observations correctement classifiées.

Résultats associés aux indicateurs de performance :

Il y a 10 indicateurs de performance affichés lorsque cette option est activée :

Exactitude, Précision, Sensibilité, F-mesure, Spécificité, Taux de Faux Positifs (TFP), Prévalence, kappa de Cohen, Taux d'erreur nul (TEN) et l'aire sous la courbe ROC (AUC).

En complément de ces indicateurs, la courbe ROC est affichée pour l'échantillon d'apprentissage et de validation (si activé). Nous avons en abscisse $1-$ la spécificité et en ordonnée la sensibilité. La courbe permet de juger visuellement le modèle, en effet, plus elle est proche du coin supérieur gauche et meilleur sera le modèle.

Il y a 2 indicateurs de performance affichés pour la régression :

la moyenne des erreurs au carré (MSE en anglais), l'erreur absolue moyenne et le coefficient de détermination R².

Les indicateurs de la première colonne correspondent à l'échantillon d'apprentissage, ceux de la seconde à l'échantillon de validation (si activé).

Résultats associés aux classes ou valeurs prédites :

Les classes ou valeurs prédites en utilisant la méthode SVM sont affichées pour les échantillons d'apprentissage, de validation et de de prédiction. De plus, dans le cas d'une classification binaire, la fonction de décision est affichée.

Résultats associés à la validation croisée :

3 indicateurs de performances sont affichés lorsque l'option de validation croisée a été cochée. Pour chaque bloc k, le taux d'erreur de classification, la F-mesure et la précision équilibrée (BAC, pour l'anglais Balanced ACcuracy), sont affichés dans le cas d'une classification binaire.

Dans le cas de la régression, la moyenne des erreurs au carré, l'erreur absolue moyenne et le coefficient de détermination R² sont affichés.

Exemple

Un tutoriel sur la façon d'utiliser les Machines à Vecteurs de Support est disponible sur le Centre d'aide XLSTAT :

http://www.xlstat.com/demo-SVMf.htm

https://help.xlstat.com/fr/6471-apprentissage-dune-regression-par-machines-vecteurs-de

Bibliographie

Vapnik, V. & Chervonenkis, A., (1964). A note on one class of perceptrons. Automation and Remote Control, 25.

Boser, B., Guyon, I. , & Vapnik, V. (1992). A training algorithm for optimal margin classifiers. In Proceedings of the Fifth Annual Workshop of Computational Learning Theory, 5, 144-152, Pittsburgh, ACM.

Cortes, C. & Vapnik V. (1995). Support-Vector Networks. Machine Learning, 20, 273-297.

Platt, J. (1998). Sequential Minimal Optimization: A fast algorithm for training support vector machines, Microsoft Research Technical Report MSR- TR-98-14.

Smola, A. & Schölkopf, B. (1998). A Tutorial on Support Vector Regression, NeuroCOLT2 Technical Report Series NC2-TR-1998-030.

Shevade, S.K., Keerthi, S.S., Bhattacharyya, C. & Murthy K.R.K. (1999). Improvements to SMO Algorithm for SVM Regression, Technical Report CD-99-16.

Fan, R., Chen, P. & Lin, C. (2005). Working Set Selection Using Second Order Information for Training Support Vector Machines, Journal of Machine Learning Research 6.

Machines à Vecteurs de Support à une classe

Utilisez cet outil pour réaliser une détection de nouveauté sur un échantillon d'observations décrit par des variables qualitatives et/ou quantitatives (prédicteurs).

Dans cette section :

Description

Les machines à vecteurs de support (SVM pour l'anglais Support Vector Machine) sont des outils d'apprentissage supervisé (Vapnik and Chervonenkis, 1964). Il fallut attendre le milieu des années 90 pour qu'une implémentation des SVM soit proposée avec l'introduction de l'astuce des noyaux (Boser, B., Guyon, I. & Vapnik V., 1992) et la généralisation au cas non séparable (Cortes C. & Vapnik V., 1995). Depuis lors, les SVM ont connu de nombreux développements et gagné en popularité dans divers domaines comme le machine learning, l'optimisation, les réseaux de neurones ou l'analyse fonctionnelle. C'est d'ailleurs l'un des algorithmes d'apprentissage qui a connu le plus de succès. Sa capacité à calculer des modèles complexes pour le coût calculatoire d'un model très simple en a fait une composante clef du domaine du machine learning où il s'est illustré en particulier dans la reconnaissance d'image ou de caractères.

Machines à Vecteurs de Support à une classe

C'est en 1999 que Schölkopf *et al.* propose une extension des SVM pour l'apprentissage non supervisé et plus précisément pour la détection de nouveauté. Nous parlons ici de détection de nouveauté, car l'algorithme apprend sur un jeu de données en partant du principe que toutes les observations sont normales. On utilise ensuite le modèle construit pour prédire si une nouvelle observation est anormale ou non.

L'algorithme des Machines à Vecteurs de Support à une classe cherche à envelopper les observations considérées "normales". L'objectif est de séparer les observations en deux classes : la classe positive considérée comme la classe des observations "normales" et la classe négative considérée comme la classe des observations "anormales". De plus, la classe positive doit contenir une grande partie des données tout en gardant une enveloppe minimale.

On veut séparer les observations par un hyperplan dont la distance à l'origine $\frac{\rho}{\|w\|}$ est maximale. Il faut alors résoudre le problème quadratique suivant :

$$\min_{w,\xi,\rho} \frac{1}{2}\|w\|^2 + \frac{1}{\nu l}\sum_{i=1}^{N}\xi_i - \rho$$

sujet à :

$$\begin{cases} (w \cdot \Phi(x_i)) \geq \rho - \xi_i, \forall i \\ \xi_i \geq 0 \end{cases}$$

Avec $\Phi : X \rightarrow F$ une transformation, X étant l'espace des observations et $\nu \in [0,1]$ le paramètre qui traduit le compromis entre le nombre d'observations appartenant à la classe positive et un $||w||^2$ minimal.

Ainsi la fonction de décision pourrait s'écrire : $$ f(x) = signe((w \cdot \Phi(x)) - \rho) $$

D'un point de vue calculatoire, le problème d'optimisation ci-dessus est quadratique avec des contraintes linéaires. Il peut donc être résolu en utilisant la méthode des multiplicateurs de Lagrange.

La fonction primale de Lagrange est :

$$L_p = \frac{1}{2}\|w\|^2 + \frac{1}{\nu l}\sum_{i=1}^{N}\xi_i - \rho - \sum_{i=1}^{N}\alpha_i[(w \cdot \phi(x_i)) - \rho + \xi_i] - \sum_{i=1}^{N}\mu_i\xi_i$$

Que l'on va minimiser par rapport à w, ξ, ρ, α et μ.

Mettre les dérivées respectives à zéro donne :

$$w = \sum_{i=1}^{N}\alpha_i\Phi(x_i)$$

$$\alpha_i = \frac{1}{\nu l} - \beta_i \leq \frac{1}{\nu l}, \sum_{i=1}^{N}\alpha_i = 1$$

La fonction duale à minimiser est :

$$L_D = \frac{1}{2}\sum_{i=1}^{N}\sum_{i'=1}^{N}\alpha_i\alpha_{i'}k(x_i, x_{i'})$$

sujette à :

$$\begin{cases} 0 \leq \alpha_i \leq \dfrac{1}{\nu N} \\ \displaystyle\sum_{i=1}^{N}\alpha_i = 1 \end{cases}$$

Avec $k(.,.)$ la fonction noyau définie positive et définie par : $k(x,y) = (\Phi(x), \Phi(y))$.

XLSTAT propose 3 noyaux en plus de la méthode linéaire :

- Noyau puissance : $k(x_i, x_i') = (\gamma \cdot (x_i^T x_i') + coefficient)^{degré}$

- Noyau RBF : $k(x_i, x_i') = e^{-\gamma \|x_i - x_i'\|^2}$
- Noyau sigmoïd : $k(x_i, x_{i'}) = \tanh(\gamma \cdot (x_i^T x_i' + coefficient))$

Enfin la fonction de décision s'écrit alors :

$$f(x) = signe(\sum_{i=1}^{N} \alpha_i k(x_i, x) - \rho)$$

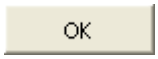
Les vecteurs de support sont identifiés comme étant alors, les observations pour lesquelles $\alpha_i > 0$.

Dans le cas ou la fonction de décision est positive ou nulle alors l'observation sera prédite normale, dans le cas contraire l'observation sera prédite anormale.

Comme pour les autres méthodes SVM présentes dans XLSTAT, l'implémentation a été possible grâce au Sequential Minimal Optimization (SMO) utilisant l'information de Second Ordre proposée par Fan *et al.* (Fan, R., Chen, P. & Lin, C., 2005).

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.


: cliquez sur ces boutons pour changer la façon dont vous chargez les données dans XLSTAT. Si le bouton possède une icône de souris, XLSTAT vous permet de faire une sélection à la souris des données. Si le bouton possède une icône de liste, XLSTAT vous permet de sélectionner les données à partir d'une liste. Si le bouton possède un icône de feuille de papier orange, des nouveaux boutons s'affichent pour charger un fichier texte ou CSV en mémoire  .

Onglet **Général** :

Variables explicatives :

- **Quantitatives** : activez cette option si vous voulez inclure une ou plusieurs variables explicatives quantitatives dans le modèle. Sélectionnez alors la ou les variables correspondantes sur la feuille Excel. Les données sélectionnées doivent être de type numérique. Si le libellé des variables a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

- **Qualitatives** : activez cette option si vous voulez inclure une ou plusieurs variables explicatives qualitatives dans le modèle. Sélectionnez alors la ou les variables correspondantes sur la feuille Excel. Les données sélectionnées peuvent être de tout type, mais les données numériques sont automatiquement considérées comme nominales. Si le libellé des variables a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

Classes connues : activez cette option si vous possédez des antécédents sur votre échantillon d'apprentissage, c'est-à-dire si vous connaissez la classe normale ou anormale de chaque observation. La sélection doit comprendre une seule variable et être binaire (1, -1, -1, 1,...). Vous devez ensuite entrer dans "Classe anormale", la valeur de la classe anormale parmi les deux classes présentes dans la sélection (exemple : "-1"). Si le libellé de la variable a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

Classe anormale : la valeur entrée correspondra à la classe des observations dites anormales.

Classe normale : dans le cas où vous ne possédez pas d'antécédents dans votre échantillon d'apprentissage, veuillez entrer une valeur qui correspondra à la classe des observations dites normales.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des variables : activez cette option si la première ligne des données sélectionnées (variables dépendantes et explicatives, libellés des observations, poids) contient un libellé.

Libellés des observations : activez cette option si vous voulez utiliser des libellés d'observations pour l'affichage des résultats. Si l'option « Libellés des variables » est activée, la première cellule de la sélection doit comprendre un en-tête. Si vous n'activez pas cette option, des libellés seront automatiquement créés (Obs1, Obs2, …).

Poids des observations : activez cette option si vous voulez utiliser des poids différents pour des observations, en particulier lorsqu'une classe est majoritairement plus présente que la seconde. Si un en-tête de colonne a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

Onglet **Options** :

Paramètres SMO : cette option vous permet de régler l'algorithme d'optimisation selon vos besoins spécifiques. Il y a 2 paramètres ajustables :

- **Nu** : ce paramètre correspond à la proportion d'observations anormales dans l'échantillon d'apprentissage. Il doit être compris entre 0 et 1.
- **Tolérance** : ce paramètre définit la tolérance appliquée lors de la comparaison de deux valeurs durant la phase d'optimisation. Ce paramètre peut être utilisé pour accélérer la vitesse de calcul.

Prétraitement : cette option vous permet de sélectionner le prétraitement appliqué aux variables explicatives. Il y a 3 options disponibles :

- **Homothétie** : les variables explicatives quantitatives sont ajustées à une échelle allant de 0 à 1 en utilisant le minimum et maximum observé pour chaque variable.
- **Normalisation** : les variables explicatives quantitatives et qualitatives sont normalisées en utilisant la moyenne et la variance observée pour chaque variable.
- **Aucun** : aucun prétraitrement n'est appliqué.

Validation croisée : disponible uniquement dans le cas où la case "Classes connues" est cochée. Cette option vous permet de lancer une validation croisée "k-fold" pour mesurer la qualité du classifieur. Les données sont divisées en k blocs de taille égales. Un seul bloc est retenu en tant qu'échantillon de validation pour tester le modèle, et les k-1 blocs restant sont utilisés en tant qu'échantillon d'apprentissage.

Noyau : cette options vous permet de sélectionner le noyau que vous souhaitez appliquer pour augmenter les dimensions de votre espace. Il y a 4 noyaux disponibles :

- **Noyau linéaire** : c'est le produit scalaire.
- **Noyau puissance** : ce noyau est détaillé dans la description. Si vous sélectionnez ce noyau, vous devez saisir la valeur du degré, du coefficient et de Gamma.
- **Noyau RBF** : c'est le noyau RBF présenté dans la description. Si vous sélectionnez ce noyau, vous devez saisir la valeur de Gamma.
- **Noyau Sigmoïde** : ce noyau est détaillé dans la description. Si vous sélectionnez ce noyau vous devez saisir la valeur du coefficient et de Gamma.

Onglet **Validation** :

Validation : disponible uniquement dans le cas où la case "Classes connues" est cochée. Activez cette option si vous souhaitez utiliser une partie des données sélectionnées pour valider le modèle.

Jeu de validation : choisissez l'une des options pour définir le mode de sélection des observations utilisées pour la validation :

- **Aléatoire** : les observations sont sélectionnées de manière aléatoire. Le « Nombre d'observations » doit alors être saisi.

- **N dernières lignes** : les N dernières observations sont sélectionnées pour la validation. Le « Nombre d'observations » N doit alors être saisi.

- **N premières lignes** : les N premières observations sont sélectionnées pour la validation. Le « Nombre d'observations » N doit alors être saisi.

- **Variable de groupe** : si vous choisissez cette option, vous devez ensuite sélectionner une variable indicatrice composée de 0 pour les observations à utiliser pour le calcul du modèle, et de 1 pour les observations à utiliser pour la validation du modèle.

Onglet **Prédiction** :

Prédiction : activez cette option si vous souhaitez sélectionner des données à utiliser en mode prédiction. Si vous activez cette option, vous devez veiller à ce que les données de prédiction soient organisées comme les données d'apprentissage : mêmes variables, même ordre dans les sélections.

Quantitatives : activez cette option pour sélectionner la ou les variables quantitatives explicatives. La première ligne doit comprendre l'en-tête si l'option libellés des variables est activée sur cet onglet.

Qualitatives : activez cette option pour sélectionner la ou les variables qualitatives explicatives. La première ligne doit comprendre l'en-tête si l'option libellés des variables est activée sur cet onglet.

Libellés des observations : activez cette option si vous voulez utiliser des libellés d'observations disponibles sur une feuille Excel pour l'affichage des résultats. Si vous n'activez pas cette option, des libellés seront automatiquement créés (PredObs1, PredObs2, …).

Libellés des variables : activez cette option si la première ligne des données sélectionnées (variables explicatives, libellés des observations) contient un libellé.

Onglet **Données manquantes** :

Ne pas accepter les valeurs manquantes : activez cette option pour que XLSTAT empêche la poursuite des calculs si des valeurs manquantes sont détectées.

Supprimer les observations : activez cette option pour supprimer les observations comportant des données manquantes.

Estimer les données manquantes : activez cette option pour estimer les données manquantes avant le début des calculs.

- **Moyenne ou mode** : activez cette option pour estimer les données manquantes en utilisant la moyenne (variables quantitatives) ou le mode (variables qualitatives) pour les variables correspondantes.

- **Plus proche voisin** : activez cette option pour estimer les données manquantes d'une observation en recherchant le plus proche voisin de l'observation.

Onglet **Sorties** :

Statistiques descriptives : activez cette option pour afficher les statistiques descriptives pour les variables sélectionnées et pour chaque jeu de données présent (apprentissage, validation, prédiction).

Résumé de l'estimation : activez cette option pour afficher un résumé du classifieur ou de la régression SVM optimisé(e).

Liste des vecteurs de support : activez cette option pour afficher la liste complète des vecteurs de support et leurs coefficients associés α, comme présentés dans la description.

Résultats par objet : activez cette option pour afficher les prédictions pour chaque observation de l'échantillon d'apprentissage, de validation et de prédiction (si activé).

Indicateurs de performance : disponible uniquement dans le cas où la case "Classes connues" est cochée. Activez cette option pour afficher la courbe ROC et les indicateurs de performance pour la classification de l'échantillon d'apprentissage et de validation (si un jeu de validation est activé).

Matrice de confusion : disponible uniquement dans le cas où la case "Classes connues" est cochée. Cette option permet d'afficher la matrice de confusion du jeu de données d'apprentissage et de validation. La matrice de confusion contient les informations concernant les classifications observées et prédites par l'algorithme. Les performances de l'algorithme peuvent être évaluées au moyen de cette matrice de confusion. La diagonale contient les prédictions correctes. Plus la somme des éléments de la diagonale est importante, meilleur est le classifieur.

Résultats

Statistiques descriptives : le tableau des statistiques descriptives présente des statistiques simples pour toutes les variables sélectionnées. Le nombre de valeurs manquantes, le nombre de valeurs non manquantes, la moyenne, l'écart standard sont affichées pour les variables quantitatives. Pour les variables qualitatives, les catégories avec leur fréquence respective et pourcentage sont affichés.

Résultat associés à l'estimation :

Un résumé descriptif de l'estimation du classifieur est affiché. La classe anormale est indiquée ainsi que la taille effective de l'échantillon d'apprentissage et les deux paramètres optimisés, le biais qui correspond au ρ et le nombre de vecteurs de support sont affichés.

Résultats associés à la liste des vecteurs de support :

Un tableau, contenant la valeur optimisée de $alpha$ et les variables explicatives prétraitrées comme elles sont utilisées durant l'optimisation, est affiché. La taille du tableau dépend du nombre de vecteurs de support identifiés.

Résultats associés aux matrices de confusion :

Les matrices de confusion sont déduites des classifications obtenues et de la classe effective ainsi que les pourcentages d'observations correctement classifiées.

Résultats associées aux indicateurs de performance :

Il y a 10 indicateurs de performance affichés lorsque cette option est activée :

Exactitude, Précision, Sensibilité, F-mesure, Spécificité, Taux de Faux Positifs (TFP), Prévalence, kappa de Cohen, Taux d'erreur nul (TEN) et l'aire sous la courbe ROC (AUC).

Les indicateurs de la première colonnes correspondent à l'échantillon d'apprentissage, ceux de la seconde à l'échantillon de validation (si activé).

En complément des ces indicateurs, la courbe ROC est affiché pour l'échantillon d'apprentissage et de validation (si activé). Nous avons en abscisse $1-$ la spécificité et en ordonnée la sensibilité. La courbe permet de juger visuellement le modèle, en effet, plus elle est proche du coin supérieur gauche et meilleur sera le modèle.

Résultats associés aux classes prédites :

Les classes prédites en utilisant la méthode SVM sont affichées pour les échantillons d'apprentissage, de validation et de de prédiction. De plus, la fonction de décision est affichée.

Résultats associés à la validation croisée :

3 indicateurs de performances sont affichés lorsque l'option de validation croisée a été cochée. Pour chaque bloc k, le taux d'erreur de classification, la F-mesure et la précision équilibrée (BAC, pour l'anglais Balanced ACcuracy) sont affichés.

Exemple

Un tutoriel sur la façon d'utiliser la SVM à une classe est disponible sur le Centre d'aide XLSTAT :

https://www.xlstat.com/demo-SVM1classfr.htm

Bibliographie

Vapnik, V. & Chervonenkis, A., (1964). A note on one class of perceptrons. Automation and Remote Control, 25.

Boser, B., Guyon, I. , & Vapnik, V. (1992). A training algorithm for optimal margin classifiers. In Proceedings of the Fifth Annual Workshop of Computational Learning Theory, 5, 144-152, Pittsburgh, ACM.

Cortes, C. & Vapnik V. (1995). Support-Vector Networks. Machine Learning, 20, 273-297.

Platt, J. (1998). Sequential Minimal Optimization: A fast algorithm for training support vector machines, Microsoft Research Technical Report MSR- TR-98-14.

Fan, R., Chen, P. & Lin, C. (2005). Working Set Selection Using Second Order Information for Training Support Vector Machines, Journal of Machine Learning Research 6.

Schölkopf, B., Williamson, R., Smola, A., Shawe-Taylor, J. & Platt, J. (1999). Support Vector Method for Novelty Detection, Microsoft NIPS. 12. 582-588.

Schölkopf, B., Platt, J., Shawe-Taylor, J., Smola, A. & Williamson, R. (2001). Estimating Support of a High-Dimensional Distribution. Neural Computation. 13. 1443-1471.

DBSCAN (Density-Based Spatial Clustering of Applications with Noise)

Utilisez cet outil pour réaliser une détection d'anomalies et une classification décrite par des variables quantitatives et/ou qualitatives.

Dans cette section :

Description

DBSCAN est l'acronyme de *Density-based Spatial Clustering of Applications with Noise* proposé par Ester, Kriegel, Sander et Xu en 1996. C'est la méthode d'apprentissage non supervisée la plus utilisée parmi les méthodes de classification par densité. Il y a plusieurs avantages au fait d'utiliser ce type de méthode, comme leur capacité à créer un nombre de classes non connu au préalable, à créer des classes non convexes et à reconnaître des anomalies.

Pour utiliser la méthode de DBSCAN, 2 paramètres sont attendus :

- $\epsilon > 0$;
- le nombre minimum de points souvent appelé $MinPts > 0$.

Plusieurs définitions permettent de comprendre la création d'une nouvelle classe. On commence par définir et compter les voisins que possèdent chaque point. On définit un **voisin** n'importe quel point p de notre jeu de données d'apprentissage qui a une distance inférieure ou égale à ϵ d'un point q.

Remarque : Par définition le point q est son propre voisin.

Chaque point peut être défini de 3 manières différentes avec l'algorithme DBSCAN :

- point central : si un point possède plus ou autant de voisins que le nombre minimum de points (entré en paramètre) alors il est considéré comme un point central ;
- point de frontière : si un point possède moins de voisins que le nombre minimum de points mais qu'il est voisin d'un point central alors il est considéré comme un point de frontière ;
- bruit : si un point n'est ni un point central ni un point de frontière, il s'agit alors d'un bruit.

On dit qu'un point p est **atteignable directement par densité** à partir d'un point q, si q est un point central et p est un voisin de q. Un point p est **atteignable par densité** lorsqu'il existe une chaîne de points qui sont chacun atteignables directement par densité à partir du point précédent. On dit aussi que deux points p et q sont **connectés par densité** s'il existe un point o à partir duquel p et q sont atteignables par densité.

Enfin, Ester *et al.* définissent une **classe** comme étant un sous-ensemble de notre jeu de données suivant deux conditions :

- dans le cas où p appartient à une classe C, si un point q est atteignable par densité à partir du point p, alors q appartient à la classe C ;
- la classe C doit respecter la condition que tous ses points doivent être mutuellement connectés par densité.

L'algorithme DBSCAN

L'algorithme DBSCAN va parcourir tous les points de notre jeu de données et va les marquer comme *visités* au fur et à mesure.

Dès qu'un premier point central est visité, une première classe est créée (nommée classe 1) ce point est alors affecté à la classe 1.

Par la suite, il va parcourir les voisins du point central en les affectant à la classe 1. Si l'algorithme trouve dans les voisins un autre point central, il va à son tour parcourir ses propres voisins et les affecter à la classe 1. Cette étape permet d'étendre la classe (*Expand Cluster*). L'algorithme arrête d'étendre la classe 1 lorsque tous les points atteignables par densité ont été parcourus.

L'algorithme continue de parcourir les points non visités du jeu de données, à la recherche d'un point central qui permettra de créer la classe 2, il pourra l'étendre si possible et ainsi de suite...

Enfin, tous les points qui ont été visités mais qui n'ont pas été affectés à une classe sont alors considérés comme des bruits.

Prédiction avec DBSCAN

La méthode de DBSCAN permet de prédire la classe de nouvelles observations.

Cela nécessite tout d'abord de trouver les voisins de chaque nouvelle observation dans le jeu de données d'apprentissage. Si la nouvelle observation possède un voisin qui a été considéré comme un point central lors de l'apprentissage, on affecte alors la nouvelle observation à la même classe que ce point central.

Si la nouvelle observation ne possède aucun point central dans ses voisins alors elle est considérée comme un bruit.

Remarque importante : l'ordre de visite des points peut influencer sur le résultat de l'affectation des points de frontière lors de l'apprentissage et de la prédiction.

L'arbre k-dimensionnel

XLSTAT vous propose d'utiliser un arbre k-dimensionnel lorsque vous ne possédez que des données quantitatives (Bentley, 1975). L'arbre va permettre de ne pas calculer toutes les distances pour trouver tous les voisins dans un rayon de taille epsilon.

L'arbre k-dimensionnel est un arbre binaire qui est construit en triant les points à partir d'une dimension et va diviser l'espace en 2 à partir de la médiane. Les points qui ont une valeur inférieure ou égale à la médiane dans cette dimension seront stockés dans le nœud fils gauche tandis que les points qui ont une valeur supérieure à la médiane seront stockés dans le nœud fils droit. La construction de l'arbre s'arrête lorsqu'il ne reste qu'un point dans un nœud.

Les formules de distance

XLSTAT propose différentes manières de calculer les distances pour pouvoir utiliser tout type de données.

Lorsque seulement des variables quantitatives sont en entrée, 5 formules de distance peuvent être utilisées :

- distance euclidienne ;
- distance de Minkowski ;
- distance de Manhattan ;
- distance de Chebychev ;
- distance de Canberra.

Lorsque seulement des variables qualitatives sont entrées, la distance d'intersection est utilisée.

Si les deux types de variables sont utilisées, la distance HEOM (Heterogeneous Euclidean Overlap Metric) est considérée.

Mesurer la performance de DBSCAN

Le coefficient de silhouette permet de mesurer le bon classement d'une observation dans sa classe. Il se calcule comme suit :

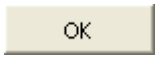
$$s(i) = \frac{b(i) - a(i)}{max(a(i), b(i))},$$

avec $a(i)$ la distance moyenne du point i par rapport aux autres points de sa classe et $b(i)$ la distance moyenne du point i par rapport aux points de la classe la plus proche.

Le coefficient de silhouette varie entre -1 et 1 et plus sa valeur est proche de 1 plus l'observation est considérée comme bien classée.

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.


: cliquez sur ce bouton pour changer la façon dont vous chargez les données dans XLSTAT. Si le bouton possède une icône de souris, XLSTAT vous permet de faire une sélection à la souris des données. Si le bouton possède une icône de liste, XLSTAT vous permet de sélectionner les données à partir d'une liste. Si le bouton possède une icône orange, des boutons complémentaires avec un point d'interrogation  sont affichés pour vous permettre d'importer les données à partir de fichiers.

Onglet **Général** :

Variables :

- **Quantitatives** : activez cette option si vous voulez inclure une ou plusieurs variables quantitatives. Sélectionnez alors la ou les variables correspondantes sur la feuille Excel. Les données sélectionnées doivent être de type numérique. Si le libellé des variables a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.
- **Qualitatives** : activez cette option si vous voulez inclure une ou plusieurs variables qualitatives. Sélectionnez alors la ou les variables correspondantes sur la feuille Excel. Les données sélectionnées peuvent être de tout type, mais les données numériques sont automatiquement considérées comme nominales. Si le libellé des variables a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

Poids des observations : activez cette option si vous voulez utiliser des poids différents pour des observations. Si un en-tête de colonne a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des variables : activez cette option si la première ligne des données sélectionnées (variables, libellés des observations, poids des observations) contient un libellé.

Libellés des observations : activez cette option si vous voulez utiliser des libellés d'observations pour l'affichage des résultats. Si l'option « Libellés des variables » est activée, la première cellule de la sélection doit comprendre un en-tête. Si vous n'activez pas cette option, des libellés seront automatiquement créés (Obs1, Obs2, …).

Onglet **Options** :

Paramètres : cette option vous permet de régler l'algorithme selon vos besoins spécifiques. Il y a 2 paramètres ajustables :

- **Epsilon** : ce paramètre correspond à la distance maximale entre deux points pour qu'ils soient considérés comme voisins. Autrement dit, si un point à une distance inférieure ou égale à ϵ d'un autre point, alors il se trouve dans le voisinage du premier point.
- **Minimum de points** : ce paramètre définit le nombre minimum de points se trouvant dans le voisinage d'un point pour qu'il soit considéré comme un point central (voir la description). Un intervalle peut être entré afin de lancer plusieurs analyses avec un nombre minimum de points différent et un pas de 1.

Prétraitement : cette option vous permet de sélectionner le prétraitement appliqué aux variables quantitatives :

- **Normalisation** : les variables quantitatives sont normalisées en utilisant la moyenne et la variance observée pour chaque variable.
- **Aucun** : aucun prétraitement n'est appliqué.

Méthode de recherche : cette option vous permet de choisir entre 2 méthodes pour permettre à l'algorithme de trouver les voisins d'un point :

- **Arbre k-dimensionnel** : vous pouvez choisir cette option seulement si des variables quantitatives sont entrées (voir description).
- **Matrice de distance** : pour tous types de variables, vous pouvez choisir d'utiliser une recherche des voisins avec la matrice de distance. Le calcul de toutes les distances permettra ensuite de calculer le coefficient de silhouette pour chaque point.

Distance : cette option vous permet de sélectionner le calcul de distance que vous souhaitez utiliser suivant le type des variables :

- pour des variables quantitatives :
- **distance euclidienne** ;
- **distance de Minkowski** ;
- **distance de Manhattan** ;
- **distance de Chebychev** ;

- **distance de Canberra**,
- pour des variables qualitatives :
    - **distance d'intersection**,
- pour des variables quantitatives et qualitatives :
    - **distance HEOM**.

Onglet **Prédiction** :

Prédiction : activez cette option si vous souhaitez sélectionner des données à utiliser en mode prédiction. Si vous activez cette option, vous devez veiller à ce que les données de prédiction soient organisées comme les données d'apprentissage : mêmes variables, même ordre dans les sélections.

Quantitatives : activez cette option pour sélectionner la ou les variables quantitatives. La première ligne doit comprendre l'en-tête si l'option libellés des variables est activée sur cet onglet.

Qualitatives : activez cette option pour sélectionner la ou les variables qualitatives. La première ligne doit comprendre l'en-tête si l'option libellés des variables est activée sur cet onglet.

Libellés des observations : activez cette option si vous voulez utiliser des libellés d'observations disponibles sur une feuille Excel pour l'affichage des résultats. Si vous n'activez pas cette option, des libellés seront automatiquement créés (PredObs1, PredObs2, …).

Libellés des variables : activez cette option si la première ligne des données sélectionnées (variables, libellés des observations) contient un libellé.

Onglet **Données manquantes** :

Ne pas accepter les valeurs manquantes : activez cette option pour que XLSTAT empêche la poursuite des calculs si des valeurs manquantes sont détectées.

Supprimer les observations : activez cette option pour supprimer les observations comportant des données manquantes.

Estimer les données manquantes : activez cette option pour estimer les données manquantes avant le début des calculs.

- **Moyenne ou mode** : activez cette option pour estimer les données manquantes en utilisant la moyenne (variables quantitatives) ou le mode (variables qualitatives) pour les variables correspondantes.
- **Plus proche voisin** : activez cette option pour estimer les données manquantes d'une observation en recherchant le plus proche voisin de l'observation.

Onglet **Sorties** :

Statistiques descriptives : activez cette option pour afficher les statistiques descriptives pour les variables sélectionnées et pour chaque jeu de données présent (apprentissage, prédiction).

Matrice de corrélation : activez cette option pour afficher un aperçu des corrélations entre les différentes variables sélectionnées pour l'échantillon d'apprentissage.

Nombre d'objets par classe : activez cette option pour afficher le nombre d'observations considérées comme des bruits et/ou le nombre d'observations attribuées à chaque classe.

Résultats par classe : activez cette option pour afficher un tableau triant les observations par classe.

Résultats par objet : activez cette option pour afficher la classe affectée à chaque observation de l'échantillon d'apprentissage. Si l'option de prédiction est activée, les classes affectées aux nouvelles observations seront toujours affichées.

- **Coefficient de silhouette** : activez cette option, disponible uniquement si vous utilisez la matrice de distance comme méthode de recherche, pour afficher les coefficients de silhouette pour chaque observation de l'échantillon d'apprentissage et le graphique associé.

Matrices de distance : activez cette option, disponible uniquement si vous utilisez la matrice de distance comme méthode de recherche, pour afficher la ou les matrices de distance.

Résultats

Statistiques descriptives : le tableau des statistiques descriptives présente des statistiques simples pour toutes les variables sélectionnées. Le nombre de valeurs manquantes, le nombre de valeurs non manquantes, la moyenne, l'écart-type sont affichés pour les variables quantitatives. Pour les variables qualitatives, les catégories avec leur fréquence respective et pourcentage sont affichées.

Matrice de corrélation : ce tableau est affiché afin de vous permettre d'avoir un aperçu des corrélations entre les différentes variables sélectionnées.

Nombre d'objets par classe : ce tableau est affiché pour avoir un aperçu de la taille de chaque classe et du nombre de bruits.

Résultats associés aux matrices de distance : une ou deux matrices de distance seront affichées si l'option de prédiction est activée. La première matrice de distance correspond aux distances entre chaque point de l'échantillon d'apprentissage. La seconde matrice de distance correspond aux distances entre les nouvelles observations et les observations de l'échantillon d'apprentissage.

Résultats associés aux objets : les classes qui ont été associées à chaque observation en utilisant l'algorithme DBSCAN sont affichées pour les échantillons d'apprentissage et de prédiction. Si la classe est 0 cela signifie que l'observation est considérée comme un bruit. De plus, le coefficient de silhouette de chaque observation est affiché dans la deuxième colonne (si l'option est activée).

Un graphique des coefficients de silhouette est affiché si l'option est activée. Les observations sont regroupées par classes dans l'ordre décroissant par rapport au coefficient de silhouette.

Résultats associés aux objets triés par classe : ce tableau affiche les observations rangées par classe.

Exemple

Un tutoriel sur la façon d'utiliser DBSCAN est disponible sur le Centre d'aide XLSTAT :

http://www.xlstat.com/demo-DBSfr.htm

Bibliographie

Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996, August). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd* (Vol. 96, No. 34, pp. 226-231).

Hahsler, M., Piekenbrock, M., & Doran, D. (2019). dbscan: Fast density-based clustering with R. *Journal of Statistical Software*, 91(1), 1-30.

Bentley, J. L. (1975). Multidimensional binary search trees used for associative searching. *Communications of the ACM*, 18(9), 509-517.

Arya, S., & Mount, D. M. (1993, March). Algorithms for fast vector quantization. In [Proceedings] *DCC93: Data Compression Conference* (pp. 381-390). IEEE.

Friedman, J. H., Bentley, J. L., & Finkel, R. A. (1977). An algorithm for finding best matches in logarithmic expected time. *ACM Transactions on Mathematical Software (TOMS)*, 3(3), 209-226.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20, 53-65.

Forêts aléatoires de classification et de régression

Utilisez cette méthode pour réaliser une classification ou une régression sur un échantillon d'observations décrites par des variables qualitatives et/ou quantitatives. La méthode permet de traiter efficacement de gros jeux de données avec un grand nombre de variables.

- **En classification** (variable réponse qualitative) : la méthode permet de prédire l'appartenance d'observations (observations, individus) à une classe d'une variable qualitative, sur la base de variables explicatives quantitatives et/ou qualitatives.
- **En régression** (variable réponse quantitative) : la méthode permet de prédire la valeur prise par une variable quantitative dépendante, en fonction de variables explicatives quantitatives et/ou qualitatives.

Dans cette section :

Description

Les forêts aléatoires sont des méthodes qui permettent d'obtenir des modèles prédictifs pour la classification et la régression. La méthode met en œuvre des arbres de décision binaire, notamment des arbres CART proposés par Breiman et al. (1984).

L'un des défauts majeurs intrinsèque de CART demeure son instabilité qui a été étudiée dans Breiman (1994). Le remède, original et profondément statistique, consiste à exploiter la variabilité naturelle des méthodes d'estimation en conjuguant deux mécanismes fondamentaux : la perturbation aléatoire des arbres et la combinaison d'un ensemble d'arbres, plutôt que la sélection de l'un d'entre eux.

L'idée générale derrière la méthode est la suivante : au lieu d'essayer d'obtenir une méthode optimisée en une fois, on génère plusieurs prédicteurs avant de mettre en commun leurs différentes prédictions.

Deux variantes sont implémentées dans XLSTAT. Le Bagging pour « Bootstrap aggregating » proposé par Breiman (1996), et la méthode Randon Input introduite dans Breiman (2001).

Le principe général de la méthode est d'agréger une collection de prédicteurs (ici des arbres CART), pour obtenir un prédicteur final plus performant.

Arbre CART

La procédure consiste à faire un partitionnement des observations en créant des groupes d'observations les plus homogènes possibles du point de vue de la variable à prédire. Plusieurs itérations sont nécessaires : à chaque itération on divise les observations en $k = 2$ classes pour expliquer la variable de sortie. La première division est obtenue en choisissant la variable explicative qui fournira la meilleure séparation des observations sur la base d'une mesure de qualité. Cette division définit des sous-populations ("nœuds") de l'arbre. L'opération est répétée pour chaque sous-population jusqu'à ce que plus aucune séparation ne soit possible. On obtient alors des nœuds terminaux appelés "feuilles" de l'arbre. Chaque feuille est caractérisée par un chemin spécifique à travers l'arbre, qu'on appelle une règle. L'ensemble des règles pour toutes les feuilles constitue le modèle.

Mesures de qualité :

- Dans le cas où la variable réponse est quantitative (régression), pour obtenir la coupure optimale à chaque nœud on cherche à minimiser la variance des nœuds fils t_L et t_R La variance d'un nœud t étant définie par :

$$\sum_{X_i \in t} (Y_i - \overline{y}(t))^2$$

avec Y_i la valeur de la variable réponse associée à l'observation i et $\overline{y}(t)$ la moyenne des sorties associées aux observations du nœud t.

- Dans le cas d'une variable réponse qualitative (classification) à J modalités, la fonction qui évalue la qualité de coupure sera ici l'indice d'impureté de Gini. L'indice d'impureté de Gini pour un nœud t est défini par :

$$i(t) = 1 - \sum_{J} p^2(j|t)$$

avec $p(j|t)$ la probabilité d'avoir la modalité J de Y sachant qu'on est dans le nœud t.

Dans le cas d'une variable explicative quantitative, tous les partitionnements binaires possible sont testés. On a donc une infinité de tests envisageables. Cependant, la taille n de l'échantillon d'apprentissage L_n étant finie, on a au plus n valeurs distinctes pour une variable continue, donc au plus $n - 1$ questions binaires associées. Pour une variable explicative qualitative, chaque regroupement en deux groupes des k modalités est testé (soit $2^k - 1$ possibilités).

Après chaque création d'un nouveau sous-nœud, les critères d'arrêt sont vérifiés, et si aucune des conditions n'est remplie, le nœud est à son tour considéré comme un nœud initial, et la procédure est itérée.

Critères d'arrêt :

Les conditions d'arrêt sont les suivantes :

- Nœud pur : le nœud ne contient que des observations correspondant à la même modalité (classification) ou à la même valeur numérique (régression).
- Variance nulle : la variance de la variable réponse associée aux observations d'un nœud est nulle.
- Aucun partitionnement ne permet d'améliorer la mesure de qualité.
- Profondeur maximale de l'arbre : le niveau du nœud correspond à la profondeur maximale de l'arbre fixée par l'utilisateur.
- Taille minimale d'un nœud parent : le nœud contient un nombre d'observations inférieur ou égal à la « taille minimale d'un nœud » fixée par l'utilisateur.
- Taille minimale d'un nœud fils : après la séparation au niveau d'un nœud, au moins l'un des sous-nœuds comprend un nombre d'observations inférieur à la « taille minimale pour un nœud fils » fixée par l'utilisateur.
- Nœuds max : le nombre maximal de nœuds terminaux a atteint la limite fixée par l'utilisateur.

Bagging

L'idée ici est qu'en construisant des arbres CART à partir de différents échantillons bootstrap, on en modifie les prédictions, et on construit ainsi une collection variée de prédicteurs. L'étape d'agrégation permet alors d'obtenir un prédicteur robuste et plus performant.

Construction :

- On construit q échantillons $L_n^1, \ldots, L_n^q$ par tirage aléatoire de n individus parmi n, ou tirage de k (défini par l'utilisateur) individus parmi n avec $k < n$ (q nombre d'arbres défini par l'utilisateur).
- Pour $l = 1, \ldots, q$ on construit l'arbre CART g_l à partir de l'échantillon L_n^l
- On agrège les prédictions des arbres $g_1, \ldots, g_q$ par :
- vote majoritaire en classification : $g(X) = max_{j \in 1 \ldots J} \sum_{l=1}^{q} I_{g_l(x)=j}$, j étant la classe prédite par l'arbre g_l pour toute observation x;
- moyenne des prédictions individuelles des arbres en régression : $\frac{1}{q} \sum_{l=1}^{q} g_l(x)$, avec $g_l(x)$ étant la valeur prédite par l'arbre g_l pour tout $x \in X$

Random Input

La variante Random Input est une modification importante du bagging, l'objectif étant de rendre les modèles (arbres) construits plus indépendants entre eux afin d'obtenir un modèle final plus

efficace. La différence fondamentale entre les deux approches est que sur chaque échantillon L_n^q on ne construit pas les arbres en suivant l'approche classique de CART, mais une variante.

Plus précisément, un arbre est ici construit de la façon suivante :

Pour découper un nœud, on tire aléatoirement un nombre m de variables ($m \leq P$, P étant le nombre de variables explicatives), et on cherche la meilleure coupure uniquement suivant les m variables sélectionnées. Le tirage, à chaque nœud, des m variables se fait sans remise et uniformément parmi toutes les variables (chaque variable a une probabilité $1/P$ d'être choisie).

Mesure d'erreur OOB :

Soit un échantillon L_n^l . Pour chaque observation (X_i, Y_i), $i = 1, \ldots, n$ de l'échantillon d'apprentissage, si L_n^l ne contient pas (X_i, Y_i) on dit que l'observation (X_i, Y_i) est « Out-Of-Bag » (OOB) pour cet échantillon.

On appellera donc échantillon OOB, l'échantillon composé de l'ensemble des observations Out-Of-Bag pour L_n^l.

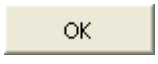
On définit l'erreur OOB comme suit :

Pour chaque observation (X_i, Y_i) $i = 1, \ldots, n$ de l'échantillon d'apprentissage L_n:

- On sélectionne les échantillons L_n^lpour lesquels (X_i, Y_i) est OOB ;
- on prédit la valeur prise par cette observation avec tous les arbres construits sur ces échantillons ;
- on agrège les prédictions de ces arbres pour fabriquer notre prédiction finale $\hat{Y}$ de$Y$;
- on calcule ensuite l'erreur commise :
  - erreur quadratique moyenne en régression : $\frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2$
  - proportion d'observations mal classées en classification : $\frac{1}{n}\sum_{i=1}^{n} I_{(\hat{Y}_i \neq Y_i)}$

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.

Onglet **Général**:

Variable réponse : Sélectionnez la variable réponse que vous souhaitez modéliser. Si des en-têtes de colonnes ont été sélectionnés, veuillez vérifier que l'option « Libellés des variables » est activée.

Type de réponse : sélectionnez le type de données correspondant à la variable réponse.

- **Qualitative** : sélectionnez cette option si vous souhaitez réaliser une classification, c'est-à-dire que la variable réponse est qualitative ou nominale.
- **Quantitative** : sélectionnez cette option si vous souhaitez réaliser une régression, c'est-à-dire que la variable réponse est quantitative.

X / Variables explicatives :

Quantitatives : activez cette option pour pouvoir sélectionner une ou plusieurs variables explicatives quantitatives. Les données sélectionnées doivent être de type numérique. Si le libellé des variables a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

Qualitatives : activez cette option pour pouvoir sélectionner une ou plusieurs variables explicatives qualitatives. Les données sélectionnées peuvent être de tout type, mais les données numériques sont automatiquement considérées comme nominales. Si le libellé des variables a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des variables : activez cette option si la première ligne des données sélectionnées (variables dépendantes et explicatives, libellés des observations) contient un libellé.

Libellés des observations : activez cette option si vous voulez utiliser des libellés d'observations pour l'affichage des résultats. Si l'option « Libellés des variables » est activée, la première cellule de la sélection doit comprendre un en-tête. Si vous n'activez pas cette option, des libellés seront automatiquement créés (Obs1, Obs2, …).

Onglet **Options**:

Paramètres de la forêt :

Échantillonnage : sélectionnez le type d'échantillons souhaité.

- **Aléatoire sans remise** : des observations sont choisies au hasard et ne peuvent figurer qu'une seule fois dans l'échantillon.
- **Aléatoire avec remise** : des observations sont choisies au hasard et peuvent figurer plusieurs fois dans l'échantillon.

Méthode : choisissez le type de forêt.

- Baging
- Random Input

Taille d'échantillon : entrez la taille k des échantillons utilisés pour la construction des arbres.

Nombre d'arbres : entrez le nombre d'arbres souhaité dans la forêt.

Conditions d'arrêt :

Temps de construction (en secondes) : entrez le temps maximal alloué à la construction de l'ensemble des arbres de la forêt. Passé ce temps, si le nombre d'arbres souhaité dans la forêt n'a pu être construit, l'algorithme s'arrête et renvoie les résultats obtenus en utilisant les arbres construits jusque-là.

Convergence : activez cette option pour vérifier la convergence de l'algorithme tous les X arbres. À partir de 100 arbres construits, tous les X arbres (X défini par l'utilisateur), on vérifie l'évolution de l'erreur OOB, et si elle est inférieure à 3%, on considère que la convergence est suffisante, et on stoppe l'algorithme.

Paramètres des arbres :

- **Taille minimale pour un parent** : entrez la taille minimale (nombre d'observations) que doit avoir un nœud parent pour être éventuellement subdivisé.
- **Taille minimale pour un fils** : entrez la taille minimale (nombre d'observations) que doit avoir un nœud fils après une subdivision pour être conservé.
- **Profondeur maximale** : entrez la profondeur maximale des arbres.

- **Nœuds max** : activez cette option pour définir le nombre maximal de nœuds terminaux que peut avoir un arbre.
- **Mtry** : nombre de variables *m* à choisir aléatoirement à chaque nœud. Notons que quand nous sommes dans le cadre du bagging.
- **Paramètre de complexité** : entrez la valeur du paramètre de complexité (CP). La construction d'un arbre ne se poursuit pas à moins de réduire l'impureté globale d'au moins un facteur CP. Cette valeur doit être inférieure à 1.

Onglet **Validation** :

Validation : activez cette option si vous souhaitez utiliser une partie des données sélectionnées pour valider le modèle.

Jeu de validation : choisissez l'une des options pour définir le mode de sélection des observations utilisées pour la validation :

- **Aléatoire** : les observations sont sélectionnées de manière aléatoire. Le « Nombre d'observations » doit alors être saisi.
- **N dernières lignes** : les N dernières observations sont sélectionnées pour la validation. Le « Nombre d'observations » N doit alors être saisi.
- **N premières lignes** : les N premières observations sont sélectionnées pour la validation. Le « Nombre d'observations » N doit alors être saisi.
- **Variable de groupe** : si vous choisissez cette option, vous devez ensuite sélectionner une variable indicatrice composée de 0 pour les observations à utiliser pour le calcul du modèle, et de 1 pour les observations à utiliser pour la validation du modèle.

Onglet **Prédiction** :

Prédiction : activez cette option si vous souhaitez sélectionner des données à utiliser en mode prédiction. Si vous activez cette option, vous devez veiller à ce que les données de prédiction soient organisées comme les données d'estimation : mêmes variables, même ordre dans les sélections, même libellés de variables si l'option est active pour les données d'apprentissage.

Quantitatives : activez cette option pour sélectionner la ou les variables quantitatives explicatives.

Qualitatives : activez cette option pour sélectionner la ou les variables qualitatives explicatives.

Libellés des observations : activez cette option si vous voulez utiliser des libellés d'observations disponibles sur une feuille Excel pour l'affichage des résultats. Si vous n'activez pas cette option, des libellés seront automatiquement créés (PredObs1, PredObs2, …).

Libellés des variables : activez cette option si la première ligne des données sélectionnées (variables explicatives, libellés des observations) contient un libellé.

Onglet **Données manquantes** :

Ne pas accepter les valeurs manquantes : activez cette option pour que XLSTAT empêche la poursuite des calculs si des valeurs manquantes sont détectées.

Supprimer les observations : activez cette option pour supprimer les observations comportant des données manquantes.

Estimer les données manquantes : activez cette option pour estimer les données manquantes avant le début des calculs.

- **Moyenne ou mode** : activez cette option pour estimer les données manquantes en utilisant la moyenne (variables quantitatives) ou le mode (variables qualitatives) pour les variables correspondantes.
- **Plus proche voisin** : activez cette option pour estimer les données manquantes d'une observation en recherchant le plus proche voisin de l'observation.

Onglet **Sorties**:

Statistiques descriptives : activez cette option pour afficher les statistiques descriptives pour les variables sélectionnées.

Prédictions OOB : activez cette option pour afficher le vecteur des prédictions Out-Of-Bag.

Détails des prédictions OOB : activez cette option pour afficher le détail des prédictions Out-Of-Bag.

Matrice de confusion (classification uniquement) : activez cette option pour afficher la matrice de confusion.

Importance des variables : activez cette option pour afficher les mesures d'importance des variables

- **Normaliser** : activez cette option pour normaliser les mesures d'importance des variables.
- **Ecart-type** : activez cette option pour afficher pour chaque variable l'écart-type de sa mesure d'importance.

Onglet **Graphiques**:

Evolution de l'erreur OOB : activez cette option pour afficher sous forme de graphique l'évolution de l'erreur OOB en fonction du nombre d'arbres.

Importance des variables : activez cette option pour afficher sous forme de graphique les mesures d'importance des variables.

Graphique des prédictions (régression uniquement) : activez cette option pour afficher le graphique des prédictions : Prédictions pour la variable dépendante versus variable

dépendante.

Résultats

Statistiques descriptives : le tableau des statistiques descriptives présente des statistiques simples pour toutes les variables sélectionnées. Le nombre de valeurs manquantes, le nombre de valeurs non manquantes, la moyenne, l'écart standard sont affichés pour les variables quantitatives. Pour les variables qualitatives, les catégories avec leur fréquence respectives et pourcentages sont affichés.

Erreur OOB :

- **Régression** : erreur quadratique moyenne (somme des carrés des résidus divisée par n)
- **Classification** : taux d'erreur de classification (basé sur les données « Out-Of-Bag »).

Graphique de l'évolution de l'erreur OOB : ce graphique montre l'évolution de l'erreur OOB en fonction du nombre d'arbres.

Matrice de confusion (uniquement en classification) : la matrice de confusion contient les informations concernant les classifications observées et prédites par l'algorithme (prédictions basées sur les données « Out-Of- Bag »). Les performances de l'algorithme peuvent être évaluées au moyen de cette matrice de confusion. La diagonale contient les prédictions correctes. Plus la somme des éléments de la diagonale est importante, meilleur est le classifieur.

Prédictions OOB : ce tableau contient différentes informations associées aux prédictions :

- **Nombre de fois OOB** : vecteur donnant pour chaque observation de l'ensemble d'apprentissage le nombre de fois qu'elle a été OOB.
- **Y et Prediction(Y)** : valeur initiale de la variable dépendante et valeur prédite (en régression on a également la valeur des résidus). La prédiction associée à chaque observation est faite en utilisant l'ensemble des arbres dans lesquels l'observation a été OOB.
- **Détails des prédictions OOB** :
    - **Régression** : tableau résumant pour chaque observation les prédictions réalisées par l'ensemble des arbres de la forêt.
    - **Classification** : tableau avec une colonne pour chaque classe de la variable réponse. Il contient pour chaque observation la probabilité qu'elle a d'appartenir aux différentes classes de la variable réponse (basé sur les données Out-Of-Bag).

Prédictions (échantillon de validation) : les classes ou valeurs prédites obtenues en utilisant le modèle prédictif construit sont affichées pour l'échantillon de validation.

Importance des variables :

La mesure d'importance calculée pour une variable donnée est l'accroissement moyen de l'erreur d'un arbre dans la forêt lorsque les valeurs observées de cette variable sont permutées au hasard dans les échantillons OOB.

Pour chaque arbre, l'erreur de prédiction sur les données out-of-bag est calculée. Ensuite, la même chose est faite après la permutation des valeurs de chaque variable explicative. La différence entre les deux est ensuite moyennée sur tous les arbres et, selon le choix de l'utilisateur, normalisée ou non par l'écart-type des différences.

Si l'écart type des différences est égal à 0 pour une variable, la division n'est pas faite.

En classification, en plus de l'impact des permutations sur l'erreur globale de la forêt, nous avons aussi l'impact sur chacune des modalités de la variable réponse.

Résultats associés aux prédictions : les classes ou valeurs prédites obtenues en utilisant le modèle prédictif construit sont affichées pour l'échantillon de prédiction.

Exemple

Des tutoriels sur la façon d'utiliser les Forêts aléatoires de classification et de régression sont disponibles sur le Centre d'aide XLSTAT :

http://www.xlstat.com/demo-rff.htm

http://www.xlstat.com/demo-rff2.htm

Bibliographie

Breiman L., Friedman J., Olshen R. and Stone C. **(1984)**. Classification And Regression Trees, Wadsworth.

Breiman L. (1996). Bagging predictors. *Machine Learning*, **24**, 123-140.

Hastie T. , Tibshirani R. and Friedman J. (2009). The Elments of Statistical Learning. Springer, Berlin.

Breiman L. (2001) Random forests.

http://www.stat.berkeley.edu/~breiman/randomforest2001.pdf

Arbres de classification et de régression

Les arbres de classification et de régression sont des méthodes qui permettent d'obtenir des modèles à la fois explicatifs et prédictifs. Parmi leurs avantages on notera d'une part leur simplicité du fait de la visualisation sous forme d'arbres, d'autre part la possibilité d'obtenir des règles en langage naturel. On distingue notamment deux cas d'utilisation de ces modèles :

- On utilise les arbres de classification pour expliquer et/ou prédire l'appartenance d'observations (observations, individus) à une classe d'une variable qualitative, sur la base de variables explicatives quantitatives et/ou qualitatives.
- On utilise les arbres de régression pour expliquer et/ou prédire la valeur prise par une variable quantitative dépendante, en fonction de variables explicatives quantitatives et/ou qualitatives.

Dans cette section :

Description

Les arbres de classification et de régression ont été proposés progressivement sous différentes formes. Les méthodes les plus utilisées sont CHAID, CART et QUEST. La méthode AID (Automatic Interaction Detection) a été proposée par Morgan et Sonquist (1963), complétée plus tard par Kass (1980) à qui l'on doit la méthode CHAID (CHi-squared Automatic Interaction Detection), puis enrichie par Biggs (1991) qui a proposé la méthode «Exhaustive CHAID». Le nom des méthodes d'arbres de classification et de régression (Classification And Regression Trees, CART) est le titre du livre introducteur de Breiman (1984). La méthode QUEST (QUick, Efficient, Statistical Tree) est plus récente (Loh et Shih, 1997).

Ces méthodes peuvent être utilisées lorsque l'on veut :

- Construire un modèle sur la base de règles, afin d'expliquer un phénomène enregistré au travers de variables dépendantes quantitatives ou qualitatives, tout en identifiant les variables explicatives les plus pertinentes.
- Identifier des groupes générés à partir des règles.
- Prévoir la valeur de la variable dépendante pour de nouvelles observations.

CHAID, CART et QUEST

XLSTAT propose quatre méthodes de construction d'arbres de classification ou de régression : CHAID, exhaustive CHAID, CART et QUEST. Dans la plupart des cas, les méthodes CHAID donnent de très bons résultats. Dans certaines situations les deux autres méthodes peuvent être intéressantes. Seul CHAID permet d'obtenir des arbres non binaires (on désigne par arbre binaire un arbre où deux branches sont créées à partir de chaque nœud).

CART

La procédure consiste à faire un partitionnement des observations en créant des groupes d'observations les plus homogènes possible du point de vue de la variable à prédire. Plusieurs itérations sont nécessaires : à chaque itération on divise les observations en $k = 2$ classes pour expliquer la variable de sortie. La première division est obtenue en choisissant la variable explicative qui fournira la meilleure séparation des observations sur la base d'une mesure de qualité. Cette division définit des sous-populations ("nœuds") de l'arbre. L'opération est répétée pour chaque sous-population jusqu'à ce que plus aucune séparation ne soit possible. On obtient alors des nœuds terminaux appelés "feuilles" de l'arbre. Chaque feuille est caractérisée par un chemin spécifique à travers l'arbre, qu'on appelle une règle. L'ensemble des règles pour toutes les feuilles constitue le modèle.

- Dans le cas où la variable dépendante est quantitative (régression), pour obtenir la coupure optimale à chaque nœud on cherche à minimiser la variance des nœuds fils (t_L et t_R) La variance d'un nœud t étant définie par :

$$\sum_{X_i \in t} (Y_i - \overline{y}(t))^2$$

avec Y_i la valeur de la variable dépendante associée à l'observation i et $\overline{y}(t)$ la moyenne des sorties associées aux observations du nœud t.

- Dans le cas d'une variable dépendante qualitative (classification) à J modalités, les fonctions évaluant la qualité de coupure seront ici l'indice d'impureté de Gini, le gain d'information (ou entropie) ou encore le critère twoing. Pour un nœud t ces mesures sont définies tel que :

$$GINI : i(t) = 1 - \sum_J p^2(j|t)$$

$$ENTROPIE : i(t) = -\sum_J p(j|t) * \log\left[p(j|t)\right]$$

$$TWOING : i(t) = \frac{p_L * p_R}{4} \left[\sum_J p^2(j|t_L) - p^2(j|t_R)\right]^2$$

avec $p(j|t)$ la probabilité d'avoir la modalité j de Y sachant qu'on est dans le nœud t, t_L et t_R respectivement les noeuds fils gauche et droit, p_L et p_R la probabilité qu'a une observation

d'appartenir resp. aux noeuds fils gauche et droit.

Dans le cas d'une variable explicative quantitative, tous les partitionnements binaires possible sont testés. On a donc une infinité de tests envisageables. Cependant, la taille n de l'échantillon d'apprentissage L_n étant finie, on a au plus n valeurs distinctes pour une variable continue, donc au plus $n-1$ questions binaires associées.

Pour une variable explicative qualitative, chaque regroupement en deux groupes des k modalités est testé (soit 2^k-1 possibilités).

Après chaque création d'un nouveau sous-nœud, les critères d'arrêt sont vérifiés, et si aucune des conditions n'est remplie, le nœud est à son tour considéré comme un nœud initial, et la procédure est itérée.

Critères d'arrêt :

Les conditions d'arrêt sont les suivantes :

- Nœud pur : le nœud ne contient que des observations correspondant à la même modalité (classification) ou à la même valeur numérique (régression).
- Variance nulle : la variance de la variable dépendante associée aux observations d'un nœud est nulle.
- Aucun partitionnement ne permet d'améliorer la mesure de qualité.
- Profondeur maximale de l'arbre : le niveau du nœud correspond à la profondeur maximale de l'arbre fixée par l'utilisateur.
- Taille minimale d'un nœud parent : le nœud contient un nombre d'observations inférieur ou égal à la « taille minimale d'un nœud parent » fixée par l'utilisateur.
- Taille minimale d'un nœud fils : après la séparation au niveau d'un nœud, au moins l'un des sous-nœuds comprend un nombre d'observations inférieur à la « taille minimale pour un nœud fils » fixée par l'utilisateur.
- Paramètre de complexité (CP) : La construction de l'arbre ne se poursuit pas à moins de réduire l'impureté globale de l'arbre d'au moins un facteur CP. Une grande valeur donne un arbre peu profond.

CHAID et Exhaustive CHAID

Ces deux méthodes procèdent en trois étapes : fusion, séparation, arrêt. Cette méthode s'appliquant uniquement sur des variables explicatives quantitatives, ces dernières sont transformées en variables qualitatives à K modalités ou catégories, $K \leq 10$.

- **Transformation des variables explicatives quantitatives:**

Soit X une variable quantitative et $a_1, a_2, \ldots, a_{K-1}$ un ensemble de points de séparation triés dans l'ordre croissant. Chaque observation x_i de X est placée dans une catégorie $C(x)$ tel que :

$$C(x) = \begin{cases} 1 & x \leq a_1 \\ \mathrm{k}+1 & a_k < x \leq a_{K-1} \\ \mathrm{K} & a_{K-1} < x \end{cases}$$

Si K est le nombre de catégories souhaité, les points de séparation sont calculés comme suit :

On calcule le rang de x_i, les poids des observations sont pris en compte dans le calcul des rangs. Dans le cas d'ex-æquos, le rang moyen est utilisé.

Soit $\left\{r_{(i)}, x_{(i)}\right\}_{i=1}^{n}$ le rang et la valeur correspondante (dans l'ordre croissant).

Pour chaque catégorie $k = 0 \ a \ (K-1), \quad I_k = \left\{i : \left\lfloor r(i) * \frac{K}{N+1} \right\rfloor = k\right\}$ où $\lfloor x \rfloor$ représente la partie entière de x. Si le groupe I_k est non vide, $i_k = max\left\{i :\in I_k\right\}$ (n étant le nombre total d'observations). Les points de séparation pour la formation des catégories sont donc les valeurs x_i correspondant aux i_k sans le plus grand.

- **Fusion :** Pour chaque variable explicative X_i, la procédure essaye de réunir les modalités similaires dans des sous-nœuds communs. Si la variable X_i ($i \leq p$ avec p le nombre total de variables) est sélectionnée pour découper le nœud. Chaque catégorie obtenue à la fin de la procédure conduira à la création d'un nœud fils. Dans cette étape sont aussi calculées les p-valeurs qui seront utilisées pour la découpe du nœud dans l'étape de séparation. Plusieurs étapes sont nécessaires :

- **Chaid :**

- Si la variable X_i ne possède que deux catégories, aller à **l'étape 7**.
- Si non, parmi les catégories de X_i, on recherche la paire de catégories la plus similaire. La paire la plus similaire étant celle qui donnera la plus grande p-valeur vis-à-vis de la variable dépendante Y. Les p-valeurs calculées le sont via un test du Khi² ou du rapport de vraisemblance en classification (si Y est qualitative) ou un test F d'ANOVA en régression (si Y est quantitative).
- Pour la paire ayant la plus grande p-valeur (noté α). Si $\alpha > \alpha_{merge}$, α_{merge} étant le seuil de regroupement défini par l'utilisateur, on procède au regroupement des deux entités. Sinon on passe à **l'étape 6**.
- (Facultatif) si la nouvelle catégorie formée rassemble plus de 2 catégories de départ, on recherche au sein de ce groupe de modalités la meilleure segmentation binaire. Ici la meilleure segmentation est celle donnant la plus faible p-valeur (α_{min}). On réalise cette segmentation si $\alpha_{min} \leq \alpha_{spli-merge}$, $\alpha_{spli-merge}$ étant le seuil de séparation défini par l'utilisateur.
- Retour à **l'étape 2**.
- Toute catégorie ayant moins d'observations que le minimum fixé par l'utilisateur est regroupée avec la catégorie la plus similaire (ayant la plus grande p-valeur).
- (Facultatif) les p-valeurs sont ajustées en appliquant la correction de Bonferonni.

- **Exhaustive CHAID:** Dans le cas de la méthode Exhaustive CHAID, cette étape utilise une méthode de recherche exhaustive pour fusionner toutes les paires similaires jusqu'à ce qu'il n'en reste plus qu'une.

- Soit $j = 0$, on calcule les p-valeurs en se basant sur l'ensemble des modalités de X_i, on la note P_j, $P_j = P_0$.

- On recherche la paire de catégorie la plus similaire (donnant la plus grande p-valeur vis-à-vis de la variable dépendante Y).
- On regroupe ces modalités dans une nouvelle entité.
- (Facultatif) si la nouvelle catégorie formée rassemble plus de 2 catégories de départ, on recherche au sein de ce groupe de modalités la meilleure segmentation binaire (donnant la plus petite p-valeur). Si cette p-valeur est supérieure à celle utilisée pour regrouper les modalités à **l'étape 3** on réalise cette segmentation.
- $j = j + 1$ puis on recalcule la p-valeur P_j en se basant sur les nouveaux groupes de X_i.
- On répète les étapes 2 à 5 jusqu'à ce qu'il ne reste que deux groupes. Ensuite parmi tous les j on cherche le regroupement donnant la plus faible p-valeur P_j.
- Toute catégorie ayant moins d'observations que le minimum fixé par l'utilisateur est regroupée avec la catégorie la plus similaire. (ayant la plus grande p-valeur).
- (Facultatif) les p-valeurs sont ajustées en appliquant la correction de Bonferonni.

Les étapes de séparation et d'arrêt sont les mêmes pour les deux méthodes.

- **Séparation :** à partir du nœud initial qui comprend la totalité des observations, la meilleure variable de séparation est celle pour laquelle la p-valeur (ou p-valeur ajustée) est la plus petite, tout en étant inférieure ou égale au « seuil de séparation » défini par l'utilisateur. Sinon la séparation n'est pas faite.

- **Arrêt :** à chaque création d'un nouveau sous-nœud, les critères d'arrêt sont vérifiés, et si aucune des conditions n'est remplie, le nœud est à son tour considéré comme un nœud parent, et la procédure est itérée. Les conditions d'arrêt sont les suivantes :

- Nœud pur : le nœud ne contient que des observations correspondant à la même modalité ou à la même valeur de la variable dépendante.

- Profondeur maximale de l'arbre : le niveau du nœud correspond à la profondeur maximale de l'arbre fixée par l'utilisateur.

- Taille minimale d'un nœud parent : le nœud contient un nombre d'observations inférieur ou égal à la « taille minimale d'un nœud » fixée par l'utilisateur.

- Taille minimale d'un nœud fils : après la séparation au niveau d'un nœud, au moins l'un des sous-nœuds comprend un nombre d'observations inférieur à la « taille minimale pour un nœud fils » fixée par l'utilisateur.

QUEST

Cette méthode ne peut être utilisée qu'avec des variables dépendantes qualitatives (classification). On procède à la séparation au niveau d'un nœud en deux sous étapes. On cherche d'abord la meilleure variable de séparation parmi les variables explicatives, puis on calcule le point de séparation pour cette variable :

- **Sélection de la variable de séparation :** pour les variables explicatives quantitatives, un test F d'ANOVA est utilisé pour comparer les moyennes correspondant aux différentes modalités de la variable dépendante Y. Pour les variables explicatives qualitatives, un

test du Khi² de Pearson est effectué. Soit X^* la variable explicative pour laquelle la p-valeur est minimale. Si cette p-valeur est inférieure à $\frac{\alpha}{p}$, où α est le seuil de signification défini par l'utilisateur et p le nombre de variables explicatives, alors X^* est choisie comme variable de séparation. Si aucune variable de séparation n'a pu être trouvée, on calcule un test de Levene pour chaque variable quantitative explicative. Soit X^{**} la variable explicative pour laquelle la p-valeur du test de Levene est minimale. Si la p-valeur est inférieure à $\frac{\alpha}{p+pX}$ où pX est le nombre de variables explicatives quantitatives, alors X^{**} est choisie comme variable de séparation. Si aucune variable de séparation n'a été trouvée, alors le nœud ne sera pas séparé en sous-nœuds.

- **Choix du point de séparation :** dans le cas d'une variable explicative qualitative, cette dernière est d'abord transformée en une variable qualitative X'. Une description détaillée de cette transformation peut être trouvée dans l'article de Loh et Shih (1997). Dans le cas d'une variable explicative quantitative, les moyennes des classes définies par les modalités de la variable dépendante sont regroupées en utilisant un algorithme k-means jusqu'à l'obtention de deux groupes. Ensuite, une analyse discriminante quadratique est réalisée sur les deux groupes afin de déterminer le point de séparation optimal.

Les probabilités à priori concernant la variable dépendante sont nécessaires pour la mise en place du modèle. XLSTAT vous donne deux possibilités pour les déterminer automatiquement à savoir : calculer les probabilités en se basant sur la répartition des classes au sein de l'ensemble d'apprentissage ou supposer que la répartition des données est la même pour chacune des modalités.

- **Arrêt :** à chaque création d'un nouveau sous-nœud, les critères d'arrêt sont vérifiés, et si aucune des conditions n'est remplie, le nœud est à son tour considéré comme un nœud parent, et la procédure est itérée. Les conditions d'arrêt sont les suivantes :
- Nœud pur : le nœud ne contient que des observations correspondant à la même modalité ou à la même valeur de la variable dépendante.
- Profondeur maximale de l'arbre : le niveau du nœud correspond à la profondeur maximale de l'arbre fixée par l'utilisateur.
- Taille minimale d'un nœud parent : le nœud contient un nombre d'observations inférieur à la « taille minimale d'un nœud » fixée par l'utilisateur.
- Taille minimale d'un nœud fils : après la séparation au niveau d'un nœud, au moins l'un des sous-nœuds comprend un nombre d'observations inférieur à la « taille minimale pour un nœud fils » fixée par l'utilisateur.

Tableau de classification et courbe ROC

Parmi les nombreux résultats proposés, XLSTAT donne la possibilité d'afficher le tableau de classification (aussi appelé matrice de confusion) qui permet de calculer un pourcentage d'observations bien classées. Lorsque seules deux classes (ou catégories, ou modalités) sont présentes dans la variable dépendante, la courbe ROC peut aussi être affichée.

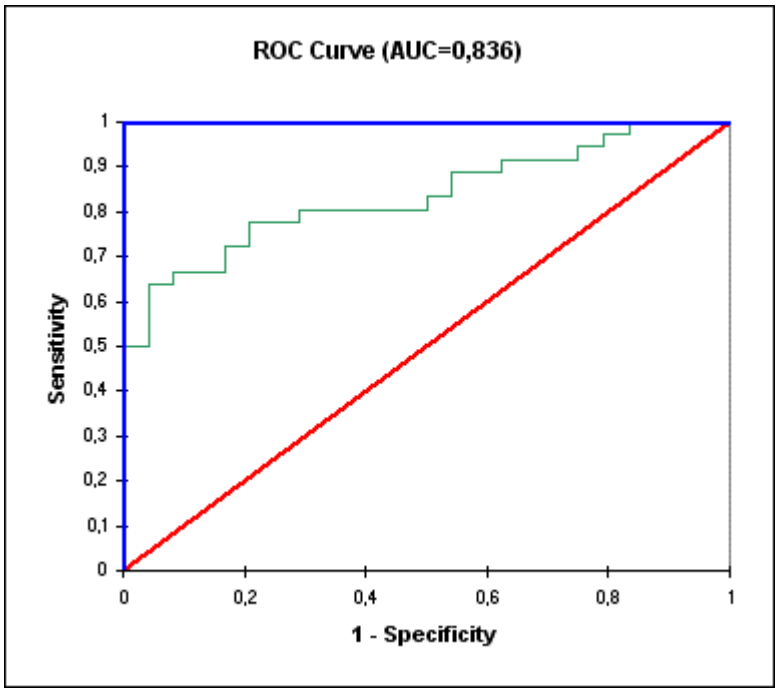
La courbe ROC (Receiver Operating Characteristics) permet de visualiser la performance d'un modèle, et de la comparer à celle d'autres modèles. Les termes utilisés viennent de la théorie de détection du signal.

On désigne par sensibilité (sensitivity) la proportion d'événements positifs bien classés. La spécificité (specificity) correspond à la proportion d'événements négatifs bien classés. Si l'on fait varier la probabilité seuil à partir de laquelle on considère qu'un événement doit être

considéré comme positif, la sensibilité et la spécificité varient. La courbe des points (1-spécificité, sensibilité) est la courbe ROC.

Considérons une variable dépendante binaire indiquant par exemple si un client a répondu favorablement à un mailing. Sur la figure ci-dessous, la courbe bleue correspond à un cas idéal où les $n\%$ de personnes ayant répondu favorablement correspondent aux $n\%$ de probabilités les plus élevées. La courbe verte correspond aux résultats d'un modèle bien discriminant.

La courbe rouge (première bissectrice) correspond à ce que l'on obtiendrait avec un modèle aléatoire de Bernoulli avec une probabilité de dépendance égale à celle observée sur l'échantillon étudié. Un modèle proche de la courbe rouge est donc inefficace puisqu'il n'est pas meilleur qu'un simple tirage aléatoire. Un modèle en dessous de cette courbe serait catastrophique car il ferait moins bien que le hasard.



L'aire sous la courbe (ou *Area Under the Curve – AUC*) est un indice synthétique calculé pour les courbes ROC. L'AUC correspond à la probabilité pour qu'un événement positif ait une probabilité donnée par le modèle plus élevée qu'un événement négatif. Pour un modèle idéal, on a AUC = 1, pour un modèle aléatoire, on a AUC = 0,5. On considère habituellement que le modèle est bon dès lors que la valeur de l'AUC est supérieure à 0,7. Un modèle bien discriminant doit avoir une AUC entre 0,87 et 0,9. Un modèle ayant une AUC supérieure à 0,9 est excellent.

Comme pour les arbres de classification, l'analyse discriminante et la régression logistique permettent de modéliser une variable qualitative. Dans le cas de variables binaires l'utilisateur pourra comparer les performances des deux méthodes en s'appuyant sur les courbes ROC.

Enfin, il est conseillé de valider le modèle sur un échantillon de validation dans la mesure du possible. XLSTAT offre plusieurs possibilités pour automatiquement générer un échantillon de validation.

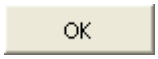
- **Courbe Lift** : la courbe Lift est la courbe qui représente la valeur Lift en fonction du pourcentage de la population. Le Lift correspond au rapport entre la proportion de vrais positifs et la proportion de prédictions positives. Un lift de 1 signifie qu'il n'existe pas de

gain par rapport à un algorithme qui ferait des prédictions de manière aléatoire. De manière générale, plus le Lift est élevé plus le modèle est performant.

- **Courbe de gain cumulée** : la courbe des gains représente la sensibilité, ou rappel, en fonction du pourcentage de population totale. Elle nous permet de voir quelle part des données concentre le maximum d'événements positifs.

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.

Onglet **Général** :

Y / Variables dépendantes : Sélectionnez la variable dépendante que vous souhaitez modéliser. Si des en-têtes de colonnes ont été sélectionnés, veuillez vérifier que l'option « Libellés des variables » est activée.

Type de réponse : sélectionnez le type de données correspondant à la variable dépendante.

- **Qualitative:** sélectionnez cette option si vous souhaitez réaliser une classification, c'est-à-dire que la variable dépendante est qualitative ou nominale.

- **Quantitative :** sélectionnez cette option si vous souhaitez réaliser une régression, c'est à dire que la variable dépendante est quantitative.

X / Variables explicatives :

- **Quantitatives :** activez cette option pour pouvoir sélectionner une ou plusieurs variables explicatives quantitatives. Les données sélectionnées doivent être de type numérique. Si le libellé des variables a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

- **Qualitatives :** activez cette option pour pouvoir sélectionner une ou plusieurs variables explicatives qualitatives. Les données sélectionnées peuvent être de tous types, mais les données numériques sont automatiquement considérées comme nominales. Si le libellé des variables a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

Poids des observations : activez cette option si vous voulez pondérer les observations. Si vous n'activez pas cette option, les poids seront tous considérés comme valant 1. XLSTAT prend en compte ces poids pour les calculs des degrés de libertés. Les poids doivent être impérativement supérieurs ou égaux à 0. Si un en-tête de colonne a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des variables : activez cette option si la première ligne des données sélectionnées (variables dépendantes et explicatives, libellés des observations, Poids des observations) contient un libellé.

Libellés des observations : activez cette option si vous voulez utiliser des libellés d'observations pour l'affichage des résultats. Si l'option « Libellés des variables » est activée, la première cellule de la sélection doit comprendre un en-tête. Si vous n'activez pas cette option, des libellés seront automatiquement créés (Obs1, Obs2, …).

Méthode : choisissez la méthode à utiliser pour les calculs parmi CHAID, Exhaustive CHAID, CART et QUEST. La méthode QUEST n'est utilisable que si la variable dépendante est qualitative.

Mesure : dans le cas des méthodes CHAID ou Exhaustive CHAID avec une variable dépendante qualitative, vous pouvez choisir d'utiliser le Khi² de Pearson et le rapport de vraisemblance. Dans le cas de la méthode CART avec une variable dépendante qualitative, vous avez le choix entre les indices de Gini, de Twoing et le gain d'information (ou entropie).

Onglet **Options** :

- Onglet **Général** :

  Paramètres de l'arbre : XLSTAT vous offre 3 possibilités pour définir les paramètres de l'arbre.

  - **Saisie manuelle :** entrez la valeur de chacun des paramètres.
  - **Automatique:** sélectionnez les paramètres que vous voulez définir de manière automatique. L'algorithme choisira alors aléatoirement la valeur de chacun des

paramètres sélectionnés, un modèle est ensuite construit avec ces paramètres et, à l'aide de la validation croisée, on évalue la qualité du modèle. Au bout de 10 itérations (donc 10 tirages aléatoires), le choix se portera sur les meilleurs paramètres (c'est à dire ceux ayant donné une erreur de validation croisée minimale) pour construire le modèle à utiliser par la suite.

- **Grille:** sélectionnez les paramètres que vous voulez définir, puis entrez les plages de valeurs à tester pour ces paramètres. L'algorithme construit l'ensemble des modèles possibles à partir de ces plages de valeurs. La meilleure combinaison de paramètres (celle avec l'erreur de validation minimale) est retenue pour construire le modèle à utiliser par la suite.

Taille minimale d'un nœud :

- **Taille minimale pour un parent :** entrez ou sélectionnez la taille minimale (nombre d'observations) que doit avoir un nœud parent pour être éventuellement subdivisé.

- **Taille minimale pour un fils :** entrez ou sélectionnez la taille minimale (nombre d'observations) que doit avoir un nœud fils après une subdivision pour être conservé.

Profondeur maximale de l'arbre : entrez ou sélectionnez la profondeur maximale de l'arbre.

Validation Croisée : (seulement en saisie manuelle) :cette option permet d'obtenir une mesure plus robuste de la qualité du modèle. La technique utilisée est la validation croisée « k-fold ». Les données sont divisées en k blocs. Parmi les k blocs, un seul bloc est retenu en tant qu'échantillon de validation pour tester le modèle, et le reste des données est utilisé en tant qu'échantillon d'apprentissage. Puis on répète l'opération en sélectionnant un autre échantillon de validation parmi les $k-1$ échantillons qui n'ont pas encore été utilisés pour la validation du modèle. L'opération se répète ainsi k fois pour qu'en fin de compte chaque sous-échantillon ait été utilisé exactement une fois comme ensemble de validation. La moyenne des k erreurs est enfin calculée pour estimer l'erreur de prédiction. Cette option est utilisée par défaut en mode **Automatique** et **Grille**, dans ces cas $k = 5$.

- **nombre de blocs :** entrez le nombre de blocs à utiliser dans la procédure de valisation croisée.

Correction du poids des classes : si les effectifs des différentes classes de la variable dépendante ne sont pas homogènes, on risque de pénaliser dans l'établissement du modèle les classes ayant un faible effectif. Afin de pallier ce problème, XLSTAT propose deux options :

- **Automatique :** le redressement est automatique. Des poids artificiels sont affectés aux observations dans le but d'obtenir des classes dont la somme des poids est identique.

- **Poids correctifs :** vous pouvez sélectionner les poids à affecter à chacune des observations.

Type de probabilité a priori (QUEST uniquement) : sélectionnez le type de probabilité à priori que vous souhaitez utiliser dans la construction de l'arbre.

Niveau de signification (%) (QUEST uniquement) : entrez le niveau de signification à utiliser pour les tests F et Khi². Des p-valeurs inférieures à cette valeur entraînent une subdivision du nœud. Cette option n'est pas disponible pour la méthode CART.

Paramètre de complexité (CP) : entrez la valeur du paramètre **CP**. La construction d'un arbre ne se poursuit pas à moins de réduire l'erreur globale d'au moins un facteur CP. Cette valeur doit être inférieure à 1.

- Onglet **CHAID** :

  Options CHAID : ces options ne sont actives qu'avec les méthodes CHAID.

  - **Seuil de séparation :** entrez la valeur du seuil de séparation. Si une p-valeur est supérieure à cette valeur, alors deux modalités ou groupes de modalités seront fusionnés.
  - **Autoriser la redivision :** activez cette option si vous voulez permettre que des modalités d'une variable qualitative explicative préalablement fusionnées, puissent être à nouveau subdivisées.
  - **Seuil de regroupement :** entrez la valeur du seuil de regroupement. Si une p-valeur est inférieure à cette valeur alors deux entités préalablement fusionnées seront divisées en deux sous groupes de catégories.
  - **Correction de Bonferroni :** activez cette option si vous souhaitez utiliser une correction de Bonferroni lors du calcul des p-valeurs associées aux modalités fusionnées.
  - **Nombre d'intervalles :** cette option n'est active que si des variables explicatives quantitatives ont été sélectionnées. Vous pouvez choisir le nombre maximum d'intervalles générés au cours de la discrétisation des variables quantitatives. Le nombre maximum d'intervalles autorisé est 10.

  Niveau de signification (%) : entrez le niveau de signification à utiliser pour les tests F et Khi². Des p-valeurs inférieures à cette valeur entraînent une subdivision du nœud. Cette option n'est pas disponible pour la méthode CART.

Onglet **Validation** :

Validation : activez cette option si vous souhaitez utiliser une partie des données sélectionnées pour valider le modèle.

Jeu de validation : choisissez l'une des options pour définir le mode de sélection des observations utilisées pour la validation :

- **Aléatoire :** les observations sont sélectionnées de manière aléatoire. Le « Nombre d'observations » doit alors être saisi.
- **N dernières lignes :** les N dernières observations sont sélectionnées pour la validation. Le « Nombre d'observations » N doit alors être saisi.

- **N premières lignes :** les N premières observations sont sélectionnées pour la validation. Le « Nombre d'observations » N doit alors être saisi.

- **Variable de groupe :** si vous choisissez cette option, vous devez ensuite sélectionner une variable indicatrice composée de 0 pour les observations à utiliser pour le calcul du modèle, et de 1 pour les observations à utiliser pour la validation du modèle.

Onglet **Prédiction** :

Prédiction : activez cette option si vous souhaitez sélectionner des données à utiliser en mode prédiction. Si vous activez cette option, vous devez veiller à ce que les données de prédiction soient organisées comme les données d'estimation : mêmes variables, même ordre dans les sélections, même libellés de variables si l'option est active pour les données d'apprentissage.

Quantitatives : activez cette option pour sélectionner la ou les variables quantitatives explicatives.

Qualitatives : activez cette option pour sélectionner la ou les variables qualitatives explicatives.

Libellés des observations : activez cette option si vous voulez utiliser des libellés d'observations disponibles sur une feuille Excel pour l'affichage des résultats. Si vous n'activez pas cette option, des libellés seront automatiquement créés (PredObs1, PredObs2, …).

Onglet **Données manquantes** :

Ne pas accepter les valeurs manquantes : activez cette option pour que XLSTAT empêche la poursuite des calculs si des valeurs manquantes sont détectées.

Supprimer les observations : activez cette option pour supprimer les observations comportant des données manquantes.

Estimer les données manquantes : activez cette option pour estimer les données manquantes avant le début des calculs.

- **Moyenne ou mode :** activez cette option pour estimer les données manquantes en utilisant la moyenne (variables quantitatives) ou le mode (variables qualitatives) pour les variables correspondantes.

- **Plus proche voisin :** activez cette option pour estimer les données manquantes d'une observation en recherchant le plus proche voisin de l'observation.

Onglet **Sorties** :

Statistiques descriptives : activez cette option pour afficher les statistiques descriptives pour les variables sélectionnées.

Corrélations : activez cette option pour afficher la matrice de corrélation pour les variables explicatives.

Structure de l'arbre : activez cette option pour afficher le tableau des nœuds, avec pour chaque nœud, le nombre d'observations, la p-valeur de la séparation, les degrès de liberté, la pureté, les nœuds-fils et le noeud parent. Dans le cas d'une variable dépendante qualitative, la modalité prédite est affichée. Pour une variable dépendante quantitative, la valeur moyenne prédite du nœud est affichée.

Fréquences des nœuds : activez cette option pour afficher le tableau des effectifs et des fréquences correspondant aux différentes modalités de la variable dépendante.

Règles : activez cette option pour afficher le tableau des règles en langage naturel correspondant aux différents nœuds et aux modalités de la variable dépendante. Seules les règles correspondant à la modalité la plus fréquente sont affichées.

Résultats par observation : activez cette option pour afficher pour chaque observation la modalité observée, la modalité prédite, et, dans le cas où la variable dépendante est qualitative, la probabilité correspondant à chacune des modalités de la variable dépendante.

Matrice de confusion (classification seulement) : activez cette option pour afficher le tableau permettant de visualiser les nombres d'observations bien et mal classées pour chacune des classes.

Onglet **Graphiques**:

Arbre : activez cette option pour afficher l'arbre de classification ou de régression.

- **Diagrammes en bâtons :** activez cette option pour afficher les nœuds sous forme d'un diagramme en bâtons, où chaque bâton correspond à une modalité de la variable dépendante.
- **Effectifs :** choisissez cette option pour afficher l'effectif correspondant à chaque barre.
- **% :** choisissez cette option pour afficher le % de la population totale correspondant à chaque barre
- **Diagrammes circulaires :** activez cette option pour afficher les nœuds sous forme d'un diagramme circulaire.

Menu contextuel pour les arbres (uniquement sous excel 2003) : Une fois l'arbre affiché, si vous cliquez sur l'un des nœuds de l'arbre, puis cliquez sur le bouton droit de votre souris, un menu contextuel est affiché avec les commandes suivantes :

Afficher tout l'arbre : cliquez sur cette commande pour afficher tout l'arbre, si vous avez déjà caché une ou plusieurs branches.

Cacher la branche : cliquez sur cette commande pour cacher les branches partant du nœud sélectionné.

Afficher la branche : cliquez sur cette commande pour afficher les branches partant du nœud sélectionné.

Définir le niveau d'élagage : choisissez cette commande pour ensuite définir un niveau d'élagage général pour l'arbre.

Réinitialiser ce menu : cliquez sur cette commande pour réinitialiser ce menu et afficher le menu Excel.

Courbe Roc : activez cette option pour afficher la courbe Roc.

Courbe Lift : activez cette option pour afficher la courbe Lift.

Courbe de gain cumulée : activez cette option pour afficher la courbe de gain cumulée.

Résultats

Statistiques descriptives : les tableaux de statistiques descriptives présentent pour toutes les variables sélectionnées des statistiques simples. pour les variables explicatives quantitatives, sont affichés le nombre d'observations, le nombre de données manquantes, le nombre de données non manquantes, la moyenne, et l'écart-type (non biaisé). Pour les variables explicatives qualitatives sont affichés le nom des différentes modalités ainsi que leur fréquence respective.

Matrice de corrélation : ce tableau est affiché afin de vous permettre d'avoir un aperçu des corrélations entre les différentes variables sélectionnées. Pour les corrélations entre variables explicatives le coefficient de corrélation de pearson est utilisé, pour celles entre variables quantitative un test du khi-deux est réalisé puis une correction sur la statistique de test via le T de Tschuprow est utilisé pour quantifier cette liaison. Ensuite pour représenter la liaison entre les variables quantitatives et qualitatives une mesure basée sur le rapport de corrélation η^2 est utilisée.

Structure de l'arbre : dans ce tableau sont affichés pour chaque nœud, le nombre d'observations, la p-valeurs de la séparation, et les deux premiers nœuds-fils. Dans le cas d'une variable dépendante qualitative, la modalité prédite est affichée. Pour une variable dépendante quantitative, la valeur moyenne prédite du nœud est affichée.

Arbre de classification ou de régression : une légende permet de repérer quel code couleur est utilisé pour chacune des modalités (variable qualitative) de la variable dépendante. La visualisation graphique de l'arbre permet de rapidement voir comment il a été itérativement construit pour aboutir à des règles d'affectation aussi pures que possible, ce qui signifie qu'idéalement les « feuilles » de l'arbre ne devraient correspondre qu'à une seule modalité.

Chaque nœud est représenté sous la forme d'un diagramme en bâtons ou d'un diagramme circulaire. Pour les diagrammes circulaires, le disque intérieur permet de visualiser la distribution des différentes modalités (ou intervalles) au niveau de ce nœud. L'anneau extérieur correspond à la distribution de ces mêmes modalités (ou intervalles au niveau du nœud parent.

L'identifiant du nœud, le nombre d'observations, l'effectif du nœud et sa pureté dans le cas d'une variable quantitative dépendante sont affichés à côté de chaque nœud. Lorsque la variable dépendante est qualitative, la prédiction est affichée à la place de la pureté. La variable de séparation est affichée entre un nœud parent et ses nœuds fils. Les flèches pointent de cette variable vers les nœuds fils. Les valeurs (des modalités pour une variable dépendante qualitative) correspondant à chacun des nœuds fils sont affichées dans le rectangle en haut à

gauche de chaque nœud fils. L'élagage de l'arbre peut être effectué grâce au menu contextuel de l'arbre (seulement sous excel 2003). Sélectionnez un nœud, puis cliquez sur le bouton droit de la souris pour afficher le menu contextuel. Les options disponibles sont décrites dans la section consacrée au menu contextuel.

Fréquence des nœuds : dans ce tableau sont fournis les effectifs et les % d'observations correspondant aux différents nœuds de l'arbre. Dans le cas où la variable dépendante est quantitative, les nombres d'observations correspondant à chacune des modalités au niveau de chaque nœud sont affichés.

Règles : dans ce tableau sont affichées les règles en langage naturel permettant d'affecter les observations à l'une ou l'autre des modalités de la variable dépendante. Ces règles, facilement compréhensibles sont facilement réutilisables. seules les règles correspondant à la modalité la plus fréquente sont affichées.

Résultats pour la recherche du meilleur modèle : ce tableau affiche tous les résultats obtenus lors de la recherche de paramètres. Sont notés en gras les paramètres retenus pour le modèle.

Résultats par objet : ce tableau indique pour chaque observation la modalité observée, la modalité prédite, et, dans le cas où la variable dépendante est qualitative, la probabilité correspondant à chacune des modalités de la variable dépendante.

Matrice de confusion (classification seulement) : ce tableau permet de visualiser les nombres d'observations bien et mal classées pour chacune des classes (voir la section description pour plus de détails).

Un tutoriel sur la façon d'utiliser les arbres de classification et de régression est disponible sur le Centre d'aide XLSTAT :

http://www.xlstat.com/demo-dtrf.htm

Bibliographie

Bigss D., Ville B. and Suen E. (1991). A method of choosing multiway partitions for classification and decision trees. *Journal of Applied Statistics*, **18(1)**, 49-62.

Goodman L. A. (1979). Simple models for the analysis of association in cross-classifications having ordered categories. *Journal of the American Statistical Association*, **74**, 537-552.

Kass G. V. (1980). An exploratory technique for investigating large quantities of categorical data. *Applied Statistics*, **20(2)**, 119-127.

Breiman L., Friedman J.H., Olshen R., and Stone C.J. (1984). Classification and Regression Tree Wadsworth & Brooks/Cole Advanced Books & Software, Pacific California.

Lim T. S., Loh W. Y. and Shih Y. S. (2000). A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms. *Machine Learning*, **40(3)**, 203-228.

Loh W. Y. and Shih Y. S., (1997). Split selection methods for classification trees. *Statistica Sinica*, **7**, 815-840.

Morgan J.N. and Sonquist J.A. (1963). Problems in the analysis of survey data and a proposal. *Journal of the American Statistical Association*, **58**, 415-434.

Rakotomalala R. (1997). Graphes d'Induction, PhD Thesis, Université Claude Bernard Lyon 1.

Rakotomalala R. (2005). TANAGRA: Une plate-forme d'expérimentation pour la fouille de données. *Revue MODULAD*, **32**, 70-85.

Bouroche J. and Tenenhaus M. (1970). Quelques méthodes de segmentation, *RAIRO*, **42**, 29-42.

Règles d'association

Utilisez ce module pour ajuster une loi de probabilité à un échantillon de données quantitatives continues ou discrètes.

Dans cette section :

Description

En 1994, Rakesh Agrawal et Ramakrishnan Sikrant ont proposés l'algorithme Apriori pour mettre en évidence des associations entre des items sous forme de règles. Cet algorithme est utilisé dans les problématiques où les volumes de données à analyser sont importants. Le nombre d'items pouvant être de plusieurs dizaines de milliers, la combinatoire est telle que toutes les règles ne peuvent être étudiées. Il convient donc de se limiter aux règles les plus importantes. Les mesures de qualités sont des valeurs probabilistes qui permettent de limiter l'explosion combinatoire pendant les 2 phases de l'algorithme et de trier les résultats.

Définitions

Items : Leur définition dépend du domaine d'application. Ils peuvent constituer des produits, des objets, des patients, des évènements, etc.

Transaction : Un ensemble d'items (minimum 1 item) étiqueté avec un identifiant unique. Les items peuvent appartenir à plusieurs transactions.

Itemset : Un groupe d'items. Les itemsets peuvent appartenir à une ou plusieurs transactions.

Support : Probabilité de retrouver un item ou un itemset X au sein d'une transaction. Elle est estimée par le nombre de fois que l'item ou l'itemset apparaît parmi toutes les transactions disponibles. Elle est comprise entre 0 et 1.

Règle : Une règle définit le lien entre deux itemsets X et Y n'ayant aucun item en commun. X->Y signifie que si X se trouve dans une transaction, Y peut apparaître dans cette même transaction.

Support d'une règle : Probabilité de trouver les items ou itemsets X et Y dans une transaction. Elle est estimée par le nombre de fois que X et Y apparaissent parmi toutes les transactions disponibles. Elle est comprise entre 0 et 1.

Confiance d'une règle : Probabilité de trouver l'item ou l'itemset Y dans une transaction, sachant que l'item ou l'itemset X se trouve dans la même transaction. Elle est estimée par la fréquence correspondante observée (nombre de fois que X et Y apparaissent parmi toutes les transactions divisé par le nombre de fois où X est trouvé). Elle est comprise entre 0 et 1.

Lift d'une règle : Le lift d'une règle est le support l'itemset regroupant X et Y divisé par le support de X et le support de Y. Le lift est symétrique (Lift(X->Y)=Lift(Y->X)) et correspond à un nombre réel positif. Un lift supérieur à 1 implique un effet positif de X sur Y (ou de Y sur X) et par, conséquent, une règle significative. Une valeur de 1 signifie qu'il n'y a pas d'effet (indépendance des items ou itemsets). Un lift inférieur à 1 signifie que X a un effet négatif sur Y ou réciproquement (exclusion d'un item ou itemset par l'autre).

Soit $I = i_1, \ldots, i_m$ un ensemble d'items. Soit $T = t_1, \ldots, t_n$ un ensemble de transactions, telles que t_i soit un sous-ensemble de I.

Une règle d'association R est représentée sous la forme suivante :

$$R : X \rightarrow Y, X \in T, Y \in T, X \cap Y = \emptyset.$$

Le support d'un sous-ensemble de I est :

$$support(X) = Pr(X).$$

La confiance d'une règle (R : X ? Y) est :

$$confiance(R) = Pr(Y|X).$$

Le lift d'une règle R : X ? Y est :

$$\text{lift}(R) = \frac{support(X \cup Y)}{support(X)support(Y)}.$$

Algorithme Apriori

Cet algorithme comporte deux étapes :

1. Génération des sous-ensembles de I qui ont un support supérieur à un support minimum fixé.

2. Génération des règles d'association à partir des sous-ensembles de I dont leur confiance est supérieure à une confiance minimum fixée.

Hiérarchie et approche multi- niveaux

XLSTAT propose de prendre en compte une hiérarchie permettant de regrouper les items et d'étudier les règles existant pour différents niveaux de groupement. La méthode proposée permet de générer des règles d'association dont les causes ou conséquences appartiennent soit à un même niveau de la hiérarchie, soit à deux niveaux différents.

Afin de simplifier la lecture des résultats, Han et Fu (1999) proposent deux valeurs alpha et beta, comprises entre 0 et 1, pour éliminer les règles redondantes et inutiles.

Une règle est dite redondante si elle est dérivée d'une règle l'englobant hiérarchiquement : la règle R telle que $A_1, ..., A_n \rightarrow B_1, ..., B_m$ est redondante, s'il existe une régle $R^{'}$ telle que $A_1^{'}, ..., A_n^{'} \rightarrow B_1^{'}, ..., B_m^{'}$ avec chacun des $A_i^{'}$ et $B_i^{'}$ $(i = 1...n)$ parent ou identique à l'un des éléments de R.

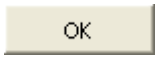
R est dite redondante si sa confiance $Conf(R)$ est comprise dans l'intervalle $[exp(Conf(R)) - \alpha, exp(Conf(R)) + \alpha]$, avec

$$exp(Conf(R)) = \frac{support(B1)}{support(B'1)} * ... * \frac{support(Bm)}{support(B'm)} * phi(R')$$

Une règle est dite inutile si elle n'apporte pas plus d'information qu'une règle avec la même conséquence ayant moins d'items comme antécédents : soient R la règle telle que (A, B -> C), et $R^{'}$ la règle (A -> C). R est considérée inutile si sa confiance Conf(R) est est dans l'intervalle $[Conf(R') - \beta, Conf(R') + \beta]$.

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT vous permet de sélectionner les données par colonnes ou par plage. Si la flèche est vers la droite, XLSTAT vous permet de sélectionner les données par lignes ou par plage.

Onglet **Général**:

Items : sélectionnez un tableau d'items et indiquez le format de données. Si des en-têtes de colonnes ont été sélectionnés, veuillez vérifier que l'option « Libellés inclus » est activée. Les

formats de données possibles sont :

- Transactionnel : choisissez ce format si vos données sont contenues dans deux colonnes, l'une indiquant la transaction (à sélectionner dans le champ Transactions), l'autre l'item. Typiquement avec ce format, on a une colonne comportant les identifiants de transaction, avec pour chaque transaction, autant de lignes qu'il y a d'items par transaction, et une colonne indiquant les items.

- Liste : choisissez ce format si vos données comprennent une ligne par transaction (les colonnes contenant les noms des items correspondant à la transaction). Le nombre d'items par transaction peut bien entendu varier d'une ligne à l'autre. Le nombre de colonnes de la sélection correspond donc au nombre maximum d'items par transaction.

- Transactions/Variables : choisissez ce format si votre tableau de données correspond à une ligne par transaction et une colonne par variable. Ce format est tel qu'il y a forcément toujours le même nombre d'item par transaction, et que les items d'une même variable ne peuvent pas être présents en même temps.

- Tableau de contingence : choisissez ce format si vos données comprennent une ligne par transaction et une colonne par item avec pour chaque transaction des valeurs nulles si l'item n'est pas présent et une valeur positive s'il est présent.

Vous avez également la possibilité de sélectionner les données dans un fichier plat en cliquant sur le bouton [...].

Transactions : sélectionnez une colonne comportant les identifiants de transaction pour chaque item. La sélection est obligatoire si le format choisi est "Transactionnel" et si le tableau d'items ne possède qu'une seule colonne. Si le tableau comprend deux colonnes, la première colonne est considérée comme correspondant à la transaction.

Items cibles / Variables cibles : Activez cette option pour définir un ou plusieurs items que vous voulez voir figurer dans la partie droite (la conséquence) des règles. Si le format de données est "Transactionnel" ou "Liste", vous devez sélectionner une liste d'items qui doivent être dans la partie droite (les conséquences) des règles qui seront générées. Si le format de données est "Transactions/Variables", vous devez sélectionner la variable qui sera considérée comme cible. Toutes les règles auront dans la partie droite (les conséquences) une modalité de la variable sélectionnée. Si le format de données est "Tableau de contingence", vous pouvez sélectionner une ou plusieurs colonnes qui seront utilisées pour identifier les items cibles.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés inclus : activez cette option si la première ligne des données sélectionnées (Items, Transactions...) contient un libellé.

Support minimal : entrez une valeur comprise entre 0 et 1 pour ne générer que les sous-ensembles d'objets ayant un support supérieur à la valeur entrée.

Confiance minimale : entrez une valeur comprise entre 0 et 1 pour ne générer que les règles ayant une confiance supérieure à la valeur entrée.

Nombre minimal d'antécédents : choisissez un nombre minimum d'antécédents aux règles générées.

Onglet **Options**:

Tri : sélectionnez une valeur sur laquelle trier les résultats (confiance, support, lift ou rien).

Onglet **Multiniveaux**:

Utiliser les informations hiérarchiques : cochez l'option si vous souhaitez sélectionner des informations hiérarchiques sur les items.

Hiérarchie : sélectionnez un tableau hiérarchique décrivant la hiérarchie entre les items et les groupes les incluant. Un item ne peut appartenir qu'à un groupe d'ordre supérieur. Vous avez la possibilité de sélectionnez les données sur un fichier plat.

Support pour chaque niveau : sélectionnez un tableau de valeurs pour affecter un support différent pour chaque niveau hiérarchique.

Analyse inter-niveaux : activez cette option si vous souhaitez générer les règles indépendamment de leur niveau hiérarchique.

Alpha (règles redondantes) : sélectionnez une valeur entre 0 et 1 pour supprimer les règles redondantes. Laissez 0 si vous ne voulez pas utiliser cette option.

Beta (règles inutiles) : sélectionnez une valeur entre 0 et 1 pour supprimer les règles inutiles. Laissez 0 si vous ne voulez pas utiliser cette option.

Onglet **Sorties**:

Matrice d'influence : activez cette option pour afficher la matrice d'influence calculée à partir de la confiance des règles d'association.

Matrice d'items : activez cette option pour visualiser l'importance relative des combinaisons entre les items.

Onglet **Graphiques** :

Graphiques d'influence : activez cette option pour afficher un graphique 2D montrant l'importance relative des diverses combinaisons obtenues par les règles d'association.

Graphiques des items : ce graphique représente en 2D l'importance relative des combinaisons entre les items.

Résultats

Règles d'association : dans ce tableau sont affichées les règles d'association obtenues par l'algorithme Apriori ainsi que les différentes valeurs relatives aux règles.

Matrice d'influence : ce tableau correspond au tableau croisé entre les antécédents et les conséquences des règles, avec pour valeur le critère choisi pour le tri des règles (confiance, support ou lift) dans l'onglet Options.

Graphique d'influence : ce graphique permet de représenter l'importance relative des règles d'association.

Matrice des items : ce tableau symmétrique indique la confiance moyenne pour chaque combinaison d'items (ligne / colonne et colonne / ligne). C'est donc un indicateur de la force de liaison entre les items. Il est ensuite utilisé pour réaliser un Multidimensional scaling (MDS) pour aboutir au **Graphiques des items** qui une représentation graphique du tableau.

Exemple

Un exemple d'utilisation de l'outil de calcul des règles d'association est disponible sur le Centre d'aide XLSTAT :

http://www.xlstat.com/demo-assocrulesf.htm

Bibliographie

Agrawal R. and Srikant R. (1994). Fast algorithms for mining association rules in large databases. In proceedings of the 20th international conference on Very Large Data Bases (VLDB'94), 487-499.

Gautam P. and Shukla R. (2012). An efficient algorithm for mining multilevel, association rule based on pincer search. Computer Application. CoRR. MANIT ,Bhopal, M.P. 462032, India.

Han J. and Fu Y. (1999). Mining multiple-level association rules in large databases. IEEE Transactions on Knowledge and Data Engineering archive, 11(5), 798-805.

Mannila H. , Toivonen H. and Inkeri Verkamo A. (1997). Discovering frequent episodes in event sequences. *Data Mining and Knowledge Discovery*, **(1)** 3, 259-289.

Indicateurs de performance de modèles

Utilisez le module Indicateurs de performance afin d'évaluer les performances de votre modèle prédictif. En fonction du type de la variable d'intérêt (quantitative ou qualitative), différents indicateurs sont proposés.

Dans cette section :

Description

Introduction

Lorsque l'on cherche à prédire les valeurs d'une variable Y de nature quantitative, on parle de **régression**. Lorsque la variable Y à prédire est de nature qualitative, on parle alors de **classification**. XLSTAT possède plusieurs modèles d'apprentissage en régression et en classification.

Nous avons donc une variable d'intérêt à prédire et plus la prédiction de l'algorithme est proche de la variable cible, plus le modèle sera performant.

Il est important de pouvoir évaluer les performances d'un modèle pour mesurer les risques mais également pour comparer plusieurs algorithmes et/ou modèles.

Le module Indicateurs de performance a été développé principalement pour nous aider à répondre à la question suivante : À quel point je peux faire confiance à un modèle pour prédire des évènements futurs ?

Indicateurs disponibles

Il existe de nombreux indicateurs pour évaluer les performances d'un modèle. Actuellement, XLSTAT propose les indicateurs suivants :

Classification

Notations : VP (Vrais Positifs), VN (Vrais Négatifs), FP (Faux Positifs) et FN (Faux Négatifs).

- **Exactitude** : l'exactitude est le rapport (VP+VN)/(VP+VN+FP+FN).
  Plus elle est proche de 1, meilleur est le test.

- **Précision** : la précision est le rapport VP/(VP + FP).
  Elle correspond à la proportion de prédictions positives effectivement correcte. En d'autres termes, un modèle ayant une précision de 0.8 prédit de manière juste la classe positive dans 80% des cas.

- **Précision équilibrée** (cas binaire uniquement) : la précision équilibrée est un indicateur utilisé pour évaluer la qualité d'un classifieur binaire. Il est particulièrement utile lorsque les classes sont déséquilibrées, c'est-à-dire que l'une des deux classes apparaît beaucoup plus souvent que l'autre. Elle est calculée comme suit : (Sensibilité + Spécificité) / 2.

- **Sensibilité** (aussi appelée **Fraction de Vrais Positifs** ou **rappel**) : proportion d'individus positifs effectivement bien détectés par le classifieur. Autrement dit, la sensibilité permet de mesurer à quel point le modèle est performant lorsqu'il est utilisé sur des individus positifs. Le modèle est parfait pour les individus positifs lorsque la sensibilité vaut 1 et est équivalent à un tirage au hasard lorsque la sensibilité vaut 0.5. Si elle est inférieure à 0.5, le modèle est contre-performant. La définition mathématique est : Sensibilité = VP/(VP + FN).

- **Spécificité** (aussi appelée **Fraction de Vrais Négatifs**) : proportion d'individus négatifs effectivement bien détectés par le test. Autrement dit, la spécificité permet de mesurer à quel point le test est performant lorsqu'il est utilisé sur des individus négatifs. Le test est parfait pour les individus négatifs lorsque la spécificité vaut 1 et est équivalent à un tirage au hasard lorsque la spécificité vaut 0.5. La définition mathématique est : Spécificité = VN/(VN + FP).

- **Fraction de faux positifs** (cas binaire uniquement) : proportion de négatifs détectés comme des positifs par le test (1-Spécificité).

- **Fraction de faux négatifs** (cas binaire uniquement) : proportion de positifs détectés comme des négatifs par le test (1-Sensibilité).

- **Bien classés** : nombre d'observations bien classées.

- **Mal classés** : nombre d'observations mal classées.

- **Prévalence de l'événement** : fréquence de survenance de l'événement dans l'échantillon total (VP+FN)/N.

- **F-mesure** : la F-mesure aussi appelée F-score ou score-F1 peut être interprétée comme une moyenne pondérée de la précision et du rappel ou sensibilité. Sa valeur est comprise entre 0 et 1. Elle est définie par : F-mesure = 2 * (Précision * Sensibilité) / (Précision + Sensibilité).

- **NER** (Taux d'erreur nul) : il correspond au pourcentage d'erreur qui serait observé si le modèle prédisait toujours la classe majoritaire.

- **Kappa de Cohen** : il est utile dans le cas où l'on veut étudier l'association entre la variable réponse et les prédictions. La valeur de Kappa est comprise entre 0 et 1 et vaut 1 lorsqu'il y a un lien total entre les deux variables (classification parfaite). Il est défini comme suit :

$$ CohenKappa = (exactitude - pe)(1 - pe) \

\text{ avec : } pe =[(VP+FN)*(VP+FP)* + *(FP+VN)*(FN+VN)] / N^2 $$

- **V de Cramer** : le test V de Cramer permet de comparer l'intensité du lien entre les deux variables étudiées. Plus V est proche de zéro, moins les variables étudiées sont dépendantes. Au contraire, il vaudra 1 lorsque les deux variables sont complètement dépendantes. Dans le cas binaire (matrice de confusion 2x2), il prend une valeur comprise entre -1 et 1. Ainsi, plus V est proche de 1, plus la liaison entre les deux variables étudiées est forte.

$$V = \sqrt{\frac{\chi^2}{W * (nClass - 1)}},$$

$nClass$ correspond au nombre de modalités de la variable réponse.

- **MCC** (coefficient de corrélation de Matthews) : le coefficient de corrélation de Matthews (MCC) ou coefficient phi est utilisé dans l'apprentissage automatique comme une mesure de la qualité des classifications binaires (à deux classes), introduit par le biochimiste Brian W. Matthews en 1975. Le MCC est défini de manière identique au coefficient phi de Pearson. Le coefficient prend en compte les vrais et faux positifs et négatifs et est généralement considéré comme une mesure équilibrée qui peut être utilisée même si les classes sont de tailles très différentes. Le MCC est essentiellement un coefficient de corrélation entre les classifications binaires observées et prédites ; il renvoie une valeur comprise entre -1 et +1. Un coefficient de 1 représente une prédiction parfaite, 0 indique que la prédiction n'est pas mieux qu'une prédiction aléatoire et -1 indique un désaccord total entre la prédiction et l'observation. (Pour plus d'informations : Coefficient de corrélation de Matthews).

  Le coefficient est calculé comme suit :

$$\frac{VP * VN - FP * FN}{\sqrt{(VP + FP) * (VP + FN) * (VN + FP) * (VN + FN)}}.$$

- **Courbe Roc** : la courbe ROC (Receiver Operating Characteristics) permet de visualiser la performance d'un modèle et de la comparer à celle d'autres modèles. Les termes utilisés viennent de la théorie de détection du signal. La courbe des points (1-spécificité, sensibilité) est la courbe ROC.

- **AUC** : l'aire sous la courbe (ou *Area Under the Curve – AUC*) est un indice synthétique calculé pour les courbes ROC. L'AUC correspond à la probabilité pour qu'un événement positif ait une probabilité donnée par le modèle plus élevée qu'un événement négatif. Pour un modèle idéal, on a AUC = 1, pour un modèle aléatoire, on a AUC = 0,5.

- **Courbe Lift** : la courbe Lift est la courbe qui représente la valeur Lift en fonction du pourcentage de la population. Le Lift correspond au rapport entre la proportion de vrais positifs et la proportion de prédictions positives. Un lift de 1 signifie qu'il n'existe pas de gain par rapport à un algorithme qui ferait des prédictions de manière aléatoire. De manière générale, plus le Lift est élevé plus le modèle est performant.

- **Courbe de gain cumulée** : la courbe des gains représente la sensibilité, ou rappel, en fonction du pourcentage de population totale. Elle nous permet de voir quelle part des données concentre le maximum d'événements positifs.

Régression

Notations :

$$\bar{y} = \frac{1}{W}\sum_{i=1}^{n} w_i y_i$$

W désigne la somme des poids.

p^* correspond au nombre de variables inclues dans le modèle.

- **MCE** : la moyenne des carrés des erreurs (MCE) est définie par :

$$MCE = \frac{1}{W}\sum_{i=1}^{n} w_i (y_i - \hat{y}_i)^2.$$

- **RMCE** : la racine de la moyenne des carrés des erreurs (RMCE) est la racine carrée de la MCE.

- **MAE** (Mean Absolute Error) :

$$MAE = \frac{1}{W}\sum_{i=1}^{n} w_i (|y_i - \hat{y}_i|).$$

- **MSLE** (Mean Squared Log Error) :

$$MSLE = \frac{1}{W}\sum_{i=1}^{n} w_i (\ln(1 + y_i) - \ln(1 + \hat{y}_i))^2.$$

- **RMSLE** (Root Mean Squared Log Error) : la RMSLE est la racine carrée de la MSLE.

- **MAPE** (Mean Absolute Percentage Error) : la MAPE aussi appelée MAPD pour Mean Absolute Percentage Deviation est définie par :

$$MAPE = \frac{1}{W}\sum_{i=1}^{n} \frac{w_i (|y_i - \hat{y}_i|)}{\max(\epsilon, |y_i|)}; \text{ avec } \epsilon = 10^{-16}.$$

  Si les valeurs observées sont très faibles ou les erreurs trop importantes, il se peut alors que le MAPE soit supérieur à 100%.

- **R²** : il correspond au coefficient de détermination du modèle. La valeur de ce coefficient est généralement comprise entre 0 et 1. XLSTAT le calcule comme suit :

$$R^2 = \frac{\sum_{i=1}^{n} w_i (\hat{y}_i - \bar{y}_i)^2}{\sum_{i=1}^{n} w_i (y_i - \bar{y}_i)^2}.$$

  Le R^2 s'interprète comme la proportion de la variabilité de la variable dépendante expliquée par le modèle. Plus le R^2 est proche de 1, meilleur est le modèle.

L'inconvénient du R^2 est qu'il ne prend pas en compte le nombre de variables utilisées pour ajuster le modèle. Sa valeur peut être négative et, dans ce cas, cela signifie que le modèle est très mal adapté aux données.

- **R² ajusté** : il correspond au coefficient de détermination ajusté du modèle. Le R^2 ajusté peut être négatif si le R^2 est voisin de zéro ou négatif. Sa valeur est définie par :

$$\hat{R}^2 = 1 - (1 - R^2)\frac{W - 1}{W - p - 1}.$$

Le R^2 ajusté est une correction du R^2 qui permet de prendre en compte le nombre de variables utilisées dans le modèle.

- **Indice de Willmott (dr)** : utilisé principalement dans les modèles hydrologiques, l'indice d'agrément redéfini (Willmott et al., 2012) est calculé comme suit :

$$

dr = \left\lbrace \begin{aligned} &1 - \frac{\sum _ { i = 1} ^ { n } w _ { i } |\hat { y } _ { i }-y_i|} {2\sum _ { i = 1} ^ { n } w _ { i }|y_i - \bar { y }| }, &si \sum _ { i = 1} ^ { n } w _ { i } |\hat { y } _ { i }-y_i| \leq 2\sum _ { i = 1} ^ { n } w _ { i }|y_i - \bar { y }| \ &\frac {2\sum _ { i = 1} ^ { n } w _ { i }|y_i - \bar { y }|}{\sum _ { i = 1} ^ { n } w _ { i } |\hat { y } _ { i }-y_i|} - 1, &si \sum _ { i = 1} ^ { n } w _ { i } |\hat { y } _ { i }-y_i| > 2\sum _ { i = 1} ^ { n } w _ { i }|y_i - \bar { y }| \end{aligned} \right.

$$

Il correspond à la somme des écarts entre les valeurs prédites par le modèle et celles observées par rapport à la somme des écarts entre le modèle parfait ($\hat{y}_i = y_i$, pour tous les i) et la moyenne observée. Ses valeurs sont comprises entre -1 et 1 et l'utilisation des valeurs absolues permet ici de limiter l'influence des valeurs extrêmes.

  - Si dr = 0 : le modèle ne fait pas mieux qu'un modèle qui prédit la moyenne observée pour toute nouvelle observation (modèle de référence). L'erreur de prédiction est la même que celle obtenue en utilisant le modèle de référence.
  - Si dr = 0.5 : la somme des erreurs de prédiction équivaut à la moitié de la somme des erreurs obtenues en utilisant le modèle de référence.
  - Si dr = -0.5 : la somme des erreurs de prédiction est égale au double de la somme des erreurs obtenues en utilisant le modèle de référence.
  - Si dr est proche de -1, cela peut indiquer que le modèle est inefficace ou que la variabilité observée ($|y_i - \bar{y}|$) est faible. De manière générale plus dr est proche de -1 plus son interprétation doit être réalisée avec précaution.

- **Indice de Mielke and Berry** : l'indice est influencé par le MAE et peut être utilisé pour les cas de saisonnalité. Il est défini comme suit :

$$ \rho = 1 -\frac{MAE}{n^{-2} \sum_{i=1}^{n}\sum_{j=1}^{} |\hat{y}_j-y_i|} $$ Le dénominateur représente la valeur moyenne de la MAE sur toutes les $n!$ permutations probables des $\hat{y}_i$

par rapport aux y_i sous l'hypothèse nulle que les n paires ($\hat{y}_i$ et y_i pour $i = 1, ...n$) correspondent au résultat d'une affectation aléatoire.

A noter que cette mesure est symétrique, c'est à dire qu'inverser $\hat{y}$ et y conduit au même résultat. Elle est bornée par 1.

- **Indice de Legates and McCabe** : utilisé principalement dans les modèles hydrologiques, l'indice de Legates and McCabe est recommandé lorsqu'il y a une saisonnalité ou une différence de moyenne par période. Il est défini comme suit :

$$E_1 = 1 - \frac{\sum_{i=1}^{n} |\hat{y}_i - y_i|}{\sum_{i=1}^{n} |y_i - \bar{y}|}$$

Une valeur de E_1 = 1 indique un modèle parfait (aucune erreur) tandis que E_1 = 0 indique un modèle qui ne fait pas mieux qu'un modèle qui prédit la moyenne observée pour toute nouvelle observation (modèle de référence). Des valeurs négatives de E_1 indiquent un modèle inefficace, car elles décrivent un "niveau d'inefficacité" par rapport au modèle de référence.

- **AIC** : le critère d'information d'Akaike (Akaike's Information Criterion) est défini par :

$$AIC = W \ln(\frac{SCE}{W}) + 2p^*$$

Ce critère proposé par Akaike (1973) dérive de la théorie de l'information et s'appuie sur la mesure de Kullback et Leibler (1951). C'est un critère de sélection de modèle qui pénalise les modèles pour lesquels l'ajout de nouvelles variables explicatives n'apporte pas suffisamment d'information au modèle (l'information est mesurée au travers de la SCE). On cherche à minimiser le critère AIC.

- **AICc** : le critère d'information d'Akaike corrigé permet de diminuer la probabilité de choisir un modèle avec un trop grand nombre de variables explicatives. Il est défini par :

$$AICc = AIC + \frac{2p^2 + 2p}{n - p - 1}$$

Ce critère serait plus performant que l'AIC lorsque le jeu de données est de petite taille et/ou possède un grand nombre de variables. Plus précisément, lorsque le rapport $\frac{n}{p}$ est inférieur à 40.

- **SBC** : le critère bayésien de Schwarz (Schwarz's Bayesian Criterion) est défini par

$$SBC = W \ln(\frac{SCE}{W}) + \ln(W)p^*$$

Ce critère proposé par Schwarz (1978) est proche du critère AIC et, comme ce dernier, on cherche à le minimiser.

Boîte de dialogue

La boîte de dialogue est composée de différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.

OK : cliquez sur ce bouton pour lancer les calculs.

Annuler : cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

Aide : cliquez sur ce bouton pour afficher l'aide.

 : cliquez sur ce bouton pour rétablir les options par défaut.

 : cliquez sur ce bouton pour effacer les sélections de données.

 : cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.

 : cliquez sur ce bouton pour changer la façon dont vous chargez les données dans XLSTAT. Si le bouton possède une icône de souris, XLSTAT vous permet de faire une sélection à la souris des données. Si le bouton possède une icône de liste, XLSTAT vous permet de sélectionner les données à partir d'une liste. Si le bouton possède une icône orange, des boutons complémentaires avec un point d'interrogation  sont affichés pour vous permettre d'importer les données à partir de fichiers.

Onglet **Général** :

Y / Variable réponse : sélectionnez la variable dépendante que vous avez modélisée. Si des en-têtes de colonnes ont été sélectionnés, veuillez vérifier que l'option « Libellés des variables » est activée.

Type de réponse : sélectionnez le type de données correspondant à la variable dépendante.

- **Qualitative :** sélectionnez cette option s'il s'agit d'une classification, c'est-à-dire que la variable dépendante est qualitative ou nominale.
- **Quantitative :** sélectionnez cette option s'il s'agit d'une régression, c'est à dire que la variable dépendante est quantitative.

Valeurs prédites : sélectionnez la/les variable(s) contenant les prédictions associées à la variable dépendante renseignée. Si des en-têtes de colonnes ont été sélectionnés, veuillez vérifier que l'option « Libellés des variables » est activée.

Si la variable réponse est qualitative, une seule variable contenant les valeurs prédites peut être renseignée.

Probabilités / Scores par classe (uniquement en classification) : sélectionnez cette option pour renseigner, pour chaque classe, les probabilités (scores) associées à chacune des observations. Si des en-têtes de colonnes ont été sélectionnés, vous devez veiller à ce que les libellés correspondent aux noms des classes auxquelles les données sont associées et que l'option « Libellés des variables » est activée. Dans le cas de classification binaire, si une seule colonne de probabilité ou de score est renseignée, alors les valeurs élevées seront associées à des prédictions de la classe positive.

Variables explicatives (uniquement en régression) : sélectionnez cette option si vous souhaitez renseigner le nombre de variables explicatives inclues dans votre modèle. Cette information est utile pour le calcul de certains indicateurs (R^2 ajusté, AIC, SBC).

Comparer à un estimateur naïf (uniquement en régression) : sélectionnez cette option si vous souhaitez comparer les performances de votre modèle avec celles obtenues en utilisant un modèle de régression naïf. Un modèle naïf désigne ici un modèle qui prédit la même valeur pour toutes les observations. XLSTAT offre 3 possibilités :

- **Moyenne** : la comparaison est effectuée avec un modèle qui prédit pour chacune des observations la moyenne de la variable dépendante.
- **Médiane** : la comparaison est effectuée avec un modèle qui prédit pour chacune des observations la médiane de la variable dépendante.
- **Définie par l'utilisateur** : la comparaison est effectuée avec un modèle qui prédit pour chacune des observations la valeur renseignée par l'utilisateur.

Afficher les résultats dans :

- **Nouvelle feuille** : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif. Dans ce cas, vous avez la possibilité de donner un nom à la feuille de résultats. Si vous n'en spécifiez pas, un nom par défaut sera créé.
- **Nouveau classeur** : activez cette option pour afficher les résultats dans un nouveau classeur. Dans ce cas, vous avez la possibilité de donner un nom à la feuille de résultats. Si vous n'en spécifiez pas, un nom par défaut sera créé.
- **Cellule existante** : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Libellés des variables : activez cette option si la première ligne des données sélectionnées contient un libellé.

Libellés des observations : activez cette option si vous voulez utiliser des libellés d'observations pour l'affichage des résultats. Si l'option « Libellés des variables » est activée, la première cellule de la sélection doit comprendre un en-tête. Si vous n'activez pas cette option, des libellés seront automatiquement créés (Obs1, Obs2, …).

Poids des observations : activez cette option si vous voulez pondérer les observations. Si vous n'activez pas cette option, les poids seront tous considérés comme valant 1. XLSTAT

prend en compte ces poids pour les calculs des degrés de libertés. Les poids doivent être impérativement supérieurs ou égaux à 0. Si un en-tête de colonne a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

Onglet **Données manquantes** :

Ne pas accepter les valeurs manquantes : activez cette option pour que XLSTAT empêche la poursuite des calculs si des valeurs manquantes sont détectées.

Supprimer les observations : activez cette option pour supprimer les observations comportant des données manquantes.

Onglet **Sorties** :

- **Classification** :

  Synthèse : activez cette option pour afficher les indicateurs suivants : Exactitude, Précision, Sensibilité, Spécificité, F-mesure, Taux de Faux Positifs (TFP), Taux de Faux Négatifs (TFN), nombre de bien classés, nombre de mal classés.

  Matrice de confusion : activez cette option pour afficher le tableau permettant de visualiser le nombre d'observations bien et mal classées pour chacune des classes.

  Prévalence : activez cette option pour calculer et afficher la prévalence.

  F-mesure : activez cette option pour calculer et afficher la F-mesure.

  Kappa de Cohen : activez cette option pour calculer et afficher le Kappa de Cohen.

  Taux d'erreur nul : activez cette option pour calculer et afficher le Taux d'erreur nul.

  Précision équilibrée : activez cette option pour calculer et afficher la précision équilibrée.

  V de Cramer : activez cette option pour calculer et afficher le V de Cramer.

  Coefficient de corrélation de Matthews : activez cette option pour calculer et afficher le Coefficient de corrélation de Matthews.

  AUC : activez cette option pour calculer et afficher la valeur de l'AUC.

- **Régression** :

  Synthèse : activez cette option pour afficher les indicateurs suivants : MAE, MCE, RMCE, MSLE, RMSLE, R^2.

  MAPE : activez cette option pour calculer et afficher la valeur du MAPE.

  Indice de Willmott : activez cette option pour calculer et afficher la valeur de l'indice de Willmott.

  Indice de Legates et McCabe : activez cette option pour calculer et afficher la valeur de l'indice de Legates et McCabe.

Indice de Mielke et Berry : activez cette option pour calculer et afficher la valeur de l'indice de Mielke et Berry.

R² ajusté : activez cette option pour calculer et afficher la valeur de l'indice du R² ajusté.

AIC de Akaike : activez cette option pour calculer et afficher la valeur de l'AIC.

SBC de Shwarz : activez cette option pour calculer et afficher la valeur du SBC.

Prédictions et résidus : activez cette option pour afficher les prédictions et les résidus pour l'ensemble des observations.

Onglet **Graphiques** :

Graphiques de régression :

- **Prédictions et résidus** : activez cette option pour afficher les graphiques suivants :
  - Variable réponse versus résidus normalisés.
  - Prédictions versus résidus normalisés.
  - Prédictions versus variable réponse.
  - Graphique en bâtons des résidus normalisés.

Graphiques de classification :

- **Courbe Roc :** activez cette option pour afficher la courbe Roc.
- **Courbe Lift :** activez cette option pour afficher la courbe Lift.
- **Courbe de gain cumulée :** activez cette option pour afficher la courbe de gain cumulée.

Résultats

Le tableau des **prédictions et résidus** permet pour chaque observation de visualiser son poids, la valeur observée de la variable dépendante, la prédiction du modèle, les résidus, les résidus standardisés et les résidus atypiques.

Les résidus atypiques sont identifiés en utilisant la méthode des boîtes à moustaches définie par TUCKEY. Dans la boîte à moustaches définie par TUKEY, la boîte a pour hauteur la distance interquartile (Q3-Q1), et les moustaches sont basées généralement sur 1,5 fois la hauteur de la boîte. Dans ce cas, une valeur est atypique si elle dépasse de 1.5 fois l'écart interquartile au-dessous du 1er quartile ou au-dessus du 3ème quartile. En se basant sur les quartiles, c'est à dire des statistiques d'ordre, la médiane et l'écart interquartile ne sont jamais influencés par les valeurs extrêmes. La valeur 1.5 est, selon TUKEY, une valeur pragmatique (rule of thumb) qui a une raison probabiliste. Si une variable suit une distribution normale, alors la zone délimitée par la boîte et les moustaches devrait contenir 99,3 % des observations. On ne devrait donc trouver que 0.7% d'observations atypiques (outliers). Si le coefficient vaut 1, la

probabilité serait de 0.957, et elle vaudrait 0.999 si le coefficient est égal à 2. Pour TUKEY la valeur 1.5 est donc un compromis pour retenir comme atypiques assez d'observations sans en retenir trop.

- Résidu(s) minimum et maximum : notés en vert (resp. rouge), ils représentent les résidus qui ont le plus petit (resp. plus grand) écart par rapport à 0. Cela nous permet aussi de voir pour chacune des observations lesquelles sont les mieux (resp. moins bien) prédites.

Exemple

Un exemple d'utilisation du module *Indicateurs de performance de modèles* est disponible sur le Centre d'aide XLSTAT à l'adresse suivante :

http://www.xlstat.com/demo-mperfs.htm

Bibliographie

Agresti A. (1990). *Categorical Data Analysis*. John Wiley and Sons, New York.

Le Guen, M. (2001). La boîte à moustaches de TUKEY, un outil pour initier à la statistique. *Statistiquement votre-SFDS, (4)*, 1-3.

Bamber D. (1975). The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of Mathematical Psychology*, **12**, 387-415.

Obuchowski, N. A. (1997). Nonparametric analysis of clustered ROC curve data. *Biometrics*, 567-578.

Baldi, P., Brunak, S., Chauvin, Y., Andersen, C. A., & Nielsen, H. (2000). Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*, 16(5), 412-424.

Willmott, C. J., Robeson, S. M., & Matsuura, K. (2012). A refined index of model performance. *International Journal of climatology*, 32(13), 2088-2094.

Labatut, V., & Cherifi, H. (2012). Accuracy Measures for the Comparison of Classifiers. The 5th International Conference on Information Technology, May 2011, amman, Jordan. pp.1,5. ffhal-00611319f

Wikipedia contributors. (2021, May 9). Matthews correlation coefficient. In Wikipedia, The Free Encyclopedia. Retrieved 10:46, May 11, 2021, from https://en.wikipedia.org/w/index.php?title=Matthews_correlation_coefficient&oldid=1022257233

Legates, D. R., & McCabe Jr, G. J. (1999). Evaluating the use of "goodness-of-fit" measures in hydrologic and hydroclimatic model validation. *Water resources research*, 35(1), 233-241.

Berry, K. J., & Mielke Jr, P. W. (1990). A generalized agreement measure. *Educational and psychological measurement*, 50(1), 123-125.

Hurvich, C. M., & Tsai, C. L. (1995). Model selection for extended quasi-likelihood models in small samples. *Biometrics*, 1077-1084.

eXtreme Gradient Boosting (XGBOOST)

XGBOOST, qui signifie "Extreme Gradient Boosting", est un modèle d'apprentissage automatique utilisé pour les problèmes d'apprentissage supervisé, dans lesquels nous utilisons un ensemble de variables explicatives pour prédire une variable cible/réponse.

XLSTAT propose une interface conviviale et sans code à la librairie populaire open-source XGBoost pour la modélisation prédictive. XGBoost est une librairie C++ scalable, portable, distribuée et open-source pour la modélisation prédictive via des arbres à gradient extrême, écrite par l'équipe dmlc ; cf. https://github.com/dmlc/xgboost

Utilisez cette méthode pour réaliser une classification ou une régression sur un échantillon d'observations décrites par des variables qualitatives et/ou quantitatives. La méthode permet de traiter efficacement de gros jeux de données avec un grand nombre de variables.

- **En classification** (variable réponse qualitative) : la méthode permet de prédire l'appartenance d'observations (observations, individus) à une classe d'une variable qualitative, sur la base de variables explicatives quantitatives et/ou qualitatives.
- **En régression** (variable réponse quantitative) : la méthode permet de prédire la valeur prise par une variable quantitative dépendante, en fonction de variables explicatives quantitatives et/ou qualitatives.

Dans cette section :

Description

Les modèles d'apprentissage automatique peuvent être adaptés aux données individuellement, ou combinés à d'autres modèles, créant ainsi un ensemble. Un ensemble est une combinaison de modèles individuels simples qui, rassemblés, créent un modèle plus performant.

En apprentissage automatique le boosting est une méthode qui permet de transformer les apprenants faibles (dans notre cas, un arbre de régression ou de classification) en apprenants forts . Il commence par construire un premier modèle sur les données puis un deuxième est ensuite construit et se concentre sur la prédiction précise des observations que le premier modèle a mal prédites. La combinaison de ces deux modèles est censée être meilleure que les modèles pris individuellement. Ce processus de boosting est ensuite répété plusieurs fois, chaque modèle successif essayant de corriger les défauts des modèles précédents.

Gradient boosting

Le boosting de gradient est un type de boosting en apprentissage automatique. Il part de l'idée que le meilleur modèle suivant possible, lorsqu'il est combiné aux modèles précédents, minimise l'erreur de prédiction globale. L'idée principale est d'établir des résultats cibles (scores/poids) pour ce prochain modèle afin de minimiser les erreurs. A chaque itération de l'algorithme, pour chaque observation, un score/poids est calculé en fonction de l'erreur de prédiction du modèle.

Le nom de boosting de gradient vient du fait que chaque poids est fixé en fonction du gradient de l'erreur relative à la prédiction. Chaque nouveau modèle fait un pas dans la direction qui minimise l'erreur de prédiction, dans l'espace des prédictions possibles pour chaque observation.

Fonction objectif: fonction de perte et regularisation

En apprentissage supervisé, la notion de modèle fait généralement référence à la structure mathématique qui permet de prédire y_i à partir de l'entrée x_i. La valeur de prédiction peut avoir différentes interprétations, en fonction de la tâche, c'est-à-dire la régression ou la classification. Par exemple, elle peut être transformée de manière logistique pour obtenir la probabilité d'une classe dans la régression logistique. Les paramètres sont la partie indéterminée que nous devons apprendre à partir des données. Nous utiliserons θ pour désigner les paramètres du modèle. L'objectif est d'entraîner le modèle pour trouver les paramètres θ qui s'adaptent le mieux les données d'apprentissage. Pour entraîner le modèle, nous devons définir la fonction objectif qui permet de mesurer l'adéquation du modèle aux données d'apprentissage.

Une caractéristique importante des fonctions objectives est qu'elles se composent de deux parties, à savoir une fonction de perte et un terme de régularisation :

$$Obj(\theta) = L(\theta) + \Omega(\theta)$$

où L est la fonction de perte et Ω le terme de régularisation. La fonction de perte mesure le pouvoir prédictif de notre modèle par rapport aux données d'apprentissage. Un choix courant pour cette fonction est l'erreur quadratique moyenne (régression).

Le terme de régularisation contrôle la complexité du modèle, ce qui nous aide à éviter le sur-apprentissage. En particulier, lorsque des observations contenant des erreurs significatives sont présentes dans les données d'entrée, l'augmentation du nombre d'itérations peut entraîner une dégradation de la performance globale plutôt qu'une amélioration.

La fonction objectif régularisée est définie :

$$Obj(\theta) = \sum_i l(\hat{y}_i, y_i) + \sum_{m=1}^{M} \Omega(\delta_m)$$
$$\text{with } \Omega(\delta) = \alpha T + \frac{1}{2}\beta \|w\|^2$$

où T est le nombre de feuilles dans l'arbre, M le nombre d'itérations, ou étapes de boosting, qui ont été effectuées. Le terme de régularisation Ω est interprété comme une combinaison de la régularisation Ridge par le coefficient β et de la pénalisation Lasso par le coefficient α. Le

terme de régularisation Ω est interprété comme une combinaison de la régularisation Ridge par le coefficient β et de la pénalisation Lasso par le coefficient α. Le terme δ_m correspond au m-ième arbre construit, le terme w correspond au vecteur de poids qui lui est attribué, et le terme w_i représente le score de la i-ième feuille.

Le principe général est que nous cherchons à avoir un modèle à la fois simple et possédant un grand pouvoir prédictif. Le compromis entre les deux est également appelé compromis biais-variance en apprentissage automatique.

Shrinkage et sous-échantillonage En plus de la fonction objective régularisée, deux techniques supplémentaires sont utilisées pour éviter le sur-apprentissage. La première technique est le shrinkage introduit par Friedman . Le shrinkage pondère les poids nouvellement ajoutés d'un facteur η après chaqueitération. Semblable à un taux d'apprentissage dans l'optimisation stochastique, le shrinkage réduit l'influence de chaque arbre individuel et laisse de la place aux arbres futurs pour améliorer le modèle.

La deuxième technique est le sous-échantillonnage des colonnes (variables). Cette technique est utilisée dans les **Forêts aléatoires** (voir la section description pour plus de détails). L'utilisation du sous-échantillonnage des colonnes permet d'éviter le sur-apprentissage encore plus que le sous-échantillonnage traditionnel des lignes (qui est également supporté).

Données manquantes

La gestion des données manquantes est traitée en interne dans l'implémentation de XGBOOST. L'algorithme propose une direction par défaut pour chaque noeud de l'arbre si une donnée est manquante. Par conséquent, le gradient est calculé uniquement sur les valeurs disponibles. Notez que les données manquantes ne sont autorisées que sur les variables explicatives x_i.

Tableau et graphiques de classification

Parmi les nombreux résultats proposés, XLSTAT donne la possibilité d'afficher le tableau de classification (aussi appelé matrice de confusion) qui permet de calculer un pourcentage d'observations bien classées. Lorsque seules deux classes (ou catégories, ou modalités) sont présentes dans la variable dépendante, la courbe ROC peut aussi être affichée.

Roc curve : La courbe ROC (Receiver Operating Characteristics) permet de visualiser la performance d'un modèle, et de la comparer à celle d'autres modèles. Les termes utilisés viennent de la théorie de détection du signal.

- **AUC**: L'aire sous la courbe (ou *Area Under the Curve – AUC*) est un indice synthétique calculé pour les courbes ROC. L'AUC correspond à la probabilité pour qu'un événement positif ait une probabilité donnée par le modèle plus élevée qu'un événement négatif. Pour un modèle idéal, on a AUC = 1, pour un modèle aléatoire, on a AUC = 0,5. On considère habituellement que le modèle est bon dès lors que la valeur de l'AUC est supérieure à 0,7. Un modèle bien discriminant doit avoir une AUC entre 0,87 et 0,9. Un modèle ayant une AUC supérieure à 0,9 est excellent.

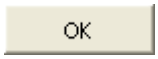
Courbe Lift : la courbe Lift est la courbe qui représente la valeur Lift en fonction du pourcentage de la population. Le Lift correspond au rapport entre la proportion de vrais positifs et la proportion de prédictions positives. Un lift de 1 signifie qu'il n'existe pas de gain par rapport à un algorithme qui ferait des prédictions de manière aléatoire. De manière générale, plus le Lift est élevé plus le modèle est performant.

Courbe de gain cumulée : la courbe des gains représente la sensibilité, ou rappel, en fonction du pourcentage de population totale. Elle nous permet de voir quelle part des données concentre le maximum d'événements positifs.

Enfin, il est conseillé de valider le modèle sur un échantillon de validation dans la mesure du possible. XLSTAT offre plusieurs possibilités pour automatiquement générer un échantillon de validation.

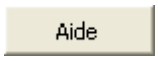
Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.



: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.


: cliquez sur ce bouton pour changer la façon dont vous chargez les données dans XLSTAT. Si le bouton possède une icône de souris, XLSTAT vous permet de faire une sélection à la souris des données. Si le bouton possède une icône de liste, XLSTAT vous permet de sélectionner les données à partir d'une liste. Si le bouton possède une icône orange, des boutons complémentaires avec un point d'interrogation  sont affichés pour vous permettre d'importer les données à partir de fichiers.

Onglet **Général** :

Y / Variable réponse : Sélectionnez la variable réponse que vous souhaitez modéliser. Si des en-têtes de colonnes ont été sélectionnés, veuillez vérifier que l'option « Libellés des variables » est activée.

Type de réponse : sélectionnez le type de données correspondant à la variable réponse.

- **Quantitative :** sélectionnez cette option si vous souhaitez réaliser une régression, c'est à dire que la variable dépendante est quantitative.

- **Binaire:** sélectionnez cette option si la variable dépendante que vous souhaitez modéliser contient exactement deux valeurs distinctes.

- **Multinomiale:** sélectionnez cette option si la variable dépendante que vous souhaitez modéliser posséde plus de 2 catégories.

X / Variables explicatives :

- **Quantitatives :** activez cette option pour pouvoir sélectionner une ou plusieurs variables explicatives quantitatives. Les données sélectionnées doivent être de type numérique. Si le libellé des variables a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

- **Qualitatives :** activez cette option pour pouvoir sélectionner une ou plusieurs variables explicatives qualitatives. Les données sélectionnées peuvent être de tous types, mais les données numériques sont automatiquement considérées comme nominales. Si le libellé des variables a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des variables : activez cette option si la première ligne des données sélectionnées (variables dépendantes et explicatives, libellés des observations, Poids des observations) contient un libellé.

Libellés des observations : activez cette option si vous voulez utiliser des libellés d'observations pour l'affichage des résultats. Si l'option « Libellés des variables » est activée, la première cellule de la sélection doit comprendre un en-tête. Si vous n'activez pas cette option, des libellés seront automatiquement créés (Obs1, Obs2, …).

Poids des observations : activez cette option si vous voulez pondérer les observations. Si vous n'activez pas cette option, les poids seront tous considérés comme valant 1. XLSTAT prend en compte ces poids pour les calculs des degrés de libertés. Les poids doivent être impérativement supérieurs ou égaux à 0. Si un en-tête de colonne a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

Onglet **Options** :

- Onglet **Général** :

Nombre maximum d'itérations : entrez le nombre maximum d'itérations de l'agorithme.

Taux d'apprentissage : entrez le paramètre de shrinkage η utilisé après chaque itération pour éviter le sur-apprentissage. Sa valeur est comprise dans l'intervalle [0,1].

Réduction minimum de l'erreur : entrez la réduction d'erreur minimale requise pour effectuer une division supplémentaire sur un noeud d'un arbre. Ses valeurs peuvent être comprises entre 0 et l'infini.

Fonction objectif : indiquez l'objectif d'apprentissage en fonction de la variable de réponse. Les options sont les suivantes :

- **Variable Reponse Quantitive**:
    - **Quadratique**: regréssion avec une fonction de perte quadratique.
    - **Log-quadratique**: régression avec une fonciton de perte log-quadratique.
    - **Logistique**: régression logistique.
    - **Pseudo-huber**: régression avec la fonction de perte Pseudo Huber, une alternative deux fois différentiable à la MAE (mean absolute error).
- **Variable réponse binaire**:
    - **Classification**: régression logistique pour la classification binaire.
- **Variable réponse multinomiale**:
    - **Classification**: régression logistique pour la classification multi-classe (fonction objectif :softmax)

Métrique : indiquez la métrique d'évaluation en fonction de l'objectif d'apprentissage. Les choix sont les suivants :

- **Variable Reponse Quantitive**:
    - **RMSE** : Root Mean Square Error.
    - **RMSLE** : Root Mean Squared Log Error.
    - **MAE** : Mean Absolute Error.
    - **MAPE** : Mean Absolute Percentage Error. La MAPE aussi appelée MAPD pour Mean Absolute Percentage Deviation
    - **MPHE** : Mean Pseudo Huber Error.
- **Variable réponse binaire**:
    - **Erreur** : taux d'erreur de classification.
    - **AUCPR** : aire sous la courbe précision / rappel
- **Variable réponse multinomiale**:
    - **Erreur** : taux d'erreur de classification.
    - **Entropie croisée** : Perte logarithmique multi-classes.

Regularisation: * **Régularisation L1**: entrez le paramètre de régularisation α. * **Régularisation L2**: entrez le paramètre de régularisation β.

Paramètres des arbres

- **Taille minimale pour un fils :** entrez la taille minimale (nombre d'observations) que doit avoir un nœud fils après une subdivision pour être conservé.

Profondeur maximale de l'arbre : entrez la profondeur maximale de l'arbre.

- Onglet **Avancées** :

Échantillonnage des lignes :

- **Taux de sous-échantillonnage** : taux de sous-échantillonnage du jeu de données d'entraînement. En le fixant à 0.5 par exemple, XGBoost replacera la moitié du jeu de données d'entraînement par un jeu de données simulé avant de générer les arbres afin de simuler le surapprentissage. Cette simulation a lieu à chaque itération.

Échantillonnage des colonnes : Il s'agit d'une famille de paramètres pour le sous-échantillonnage des colonnes. Les paramètres peuvent prendre des valeurs dans l'intervalle (0, 1). Ils ont une valeur par défaut de 1, et spécifient la proportion de colonnes à sous-échantillonner.

- **Par arbre** : taux de sous-échantillonnage des colonnes lors de la construction de chaque arbre. Le sous-échantillonnage a lieu une fois pour chaque arbre construit.

- **Par niveau** : taux de sous-échantillonnage des colonnes pour chaque niveau. Le sous-échantillonnage a lieu une fois pour chaque nouveau niveau de profondeur atteint dans un arbre.

- **Par noeud** : taux de sous-échantillonnage des colonnes pour chaque nœud (division). Le sous-échantillonnage a lieu une fois à chaque fois qu'une nouvelle division est évaluée. Les colonnes sont sous-échantillonnées à partir de l'ensemble des colonnes choisies pour le niveau actuel.

Onglet **Validation** :

Validation : activez cette option si vous souhaitez utiliser une partie des données sélectionnées pour valider le modèle.

Jeu de validation : choisissez l'une des options pour définir le mode de sélection des observations utilisées pour la validation :

- **Aléatoire :** les observations sont sélectionnées de manière aléatoire. Le « Nombre d'observations » doit alors être saisi.

- **N dernières lignes :** les N dernières observations sont sélectionnées pour la validation. Le « Nombre d'observations » N doit alors être saisi.

- **N premières lignes :** les N premières observations sont sélectionnées pour la validation. Le « Nombre d'observations » N doit alors être saisi.

- **Variable de groupe :** si vous choisissez cette option, vous devez ensuite sélectionner une variable indicatrice composée de 0 pour les observations à utiliser pour le calcul du modèle, et de 1 pour les observations à utiliser pour la validation du modèle.

Onglet **Prédiction** :

Prédiction : activez cette option si vous souhaitez sélectionner des données à utiliser en mode prédiction. Si vous activez cette option, vous devez veiller à ce que les données de prédiction

soient organisées comme les données d'estimation : mêmes variables, même ordre dans les sélections, même libellés de variables si l'option est active pour les données d'apprentissage.

Quantitatives : activez cette option pour sélectionner la ou les variables quantitatives explicatives.

Qualitatives : activez cette option pour sélectionner la ou les variables qualitatives explicatives.

Libellés des observations : activez cette option si vous voulez utiliser des libellés d'observations disponibles sur une feuille Excel pour l'affichage des résultats. Si vous n'activez pas cette option, des libellés seront automatiquement créés (PredObs1, PredObs2, …).

Libellés des variables : activez cette option si la première ligne des données sélectionnées (variables dépendantes et explicatives, libellés des observations, Poids des observations) contient un libellé.

Onglet **Données manquantes** :

Ne pas accepter les valeurs manquantes : activez cette option pour que XLSTAT empêche la poursuite des calculs si des valeurs manquantes sont détectées.

Supprimer les observations : activez cette option pour supprimer les observations comportant des données manquantes.

Ignorer les données manquantes : Activez cette option pour ignorer les données manquantes. Si des données manquantes sont présentes pour la ou les variables explicatives, elles seront traitées par l'algorithme XGBOOST en utilisant le processus décrit dans la description. Les données manquantes ne sont pas autorisées dans la variable de réponse.

Estimer les données manquantes : activez cette option pour estimer les données manquantes avant le début des calculs.

- **Moyenne ou mode :** activez cette option pour estimer les données manquantes en utilisant la moyenne (variables quantitatives) ou le mode (variables qualitatives) pour les variables correspondantes.
- **Plus proche voisin :** activez cette option pour estimer les données manquantes d'une observation en recherchant le plus proche voisin de l'observation.

Onglet **Sorties** :

Statistiques descriptives : activez cette option pour afficher les statistiques descriptives pour les variables sélectionnées.

Corrélations : activez cette option pour afficher la matrice de corrélation pour les variables explicatives.

Prédictions et résidus (régression seulement): activez cette option pour afficher les prédictions et les résidus pour l'ensemble des observations.

Résultats par objet (classification seulement) : activez cette option pour afficher pour chaque observation la modalité observée, la modalité prédite, et, dans le cas où la variable dépendante

est qualitative, la probabilité correspondant à chacune des modalités de la variable dépendante.

Statistiques pour chaque itération : Activez cette option pour afficher le tableau montrant l'évolution des mesures d'évaluation pour chaque itération.

Matrice de confusion (classification seulement) : activez cette option pour afficher le tableau permettant de visualiser les nombres d'observations bien et mal classées pour chacune des classes.

Importance des variables : Activez cette option pour afficher les mesures d'importance des variables. XLSTAT affiches les mesures d'importance suivantes :

La **Fréquence** correspond au nombre de fois qu'une variable est utilisée pour diviser les données sur l'ensemble des arbres du modèle. il est exprimé en pourcentage de l'ensemble des divisions réalisées.

Le **Gain** correspond à la contribution relative d'une variable au modèle et est calculé en prenant le rapport entre la contribution totale de la variable considérée et celle de l'ensemble des variables du modèle. La valeur de cette mesure est proportionnelle à l'implication de la variable pour générer une prédiction.

La **Couverture** correspond à la proportion d'observations liées à une variable. Lorsqu'une variable est utilisée pour diviser un nœud qui précède une feuille, on dit alors que les observations présentes dans ce nœud sont couvertes par la variable. Par exemple, vous possédez un jeu de données avec 100 observations et 4 variables, 3 arbres sont construits. La variable 1 est utilisée pour diviser un nœud précédant une feuille. Ces nœuds comportent respectivement 10, 5 et 2 observations dans les arbres 1, 2 et 3 respectivement. La couverture pour la variable 1 correspond à la somme des observations : 10+5+2 = 17 observations. Ce nombre est ensuite exprimé en pourcentage de la mesure de couverture sur l'ensemble des arbres du modèle.

Le gain est l'attribut le plus pertinent pour interpréter l'importance relative de chaque variable.

Onglet **Graphiques**:

Statistiques pour chaque itération : activez cette option pour afficher sous forme graphique l'évolution de la métrique d'évaluation en fonction du nombre d'itération.

Importance des variables : activez cette option pour afficher sous forme de graphique les mesures d'importance des variables.

Graphiques de régression :

- Variable réponse versus résidus normalisés.
- Prédictions versus résidus normalisés.
- Prédictions versus variable réponse.
- Graphique en bâtons des résidus normalisés.

Graphique de confusion (classification seulement): activez cette option pour afficher le graphique de confusion qui permet une visualisation synthétique du tableau de classification. Les effectifs peuvent être liés soit à la largeur, soit l'aire, des carrés représentés.

Courbe Roc (classification seulement): activez cette option pour afficher la courbe Roc.

Courbe Lift(classification seulement): activez cette option pour afficher la courbe Lift.

Courbe de gain cumulée(classification seulement): activez cette option pour afficher la courbe de gain cumulée.

Résultats

Le tableau des **prédictions et résidus** permet pour chaque observation de visualiser son poids, la valeur observée de la variable dépendante, la prédiction du modèle, les résidus, les résidus standardisés et les résidus atypiques.

Les résidus atypiques sont identifiés en utilisant la méthode des boîtes à moustaches définie par TUCKEY. Dans la boîte à moustaches définie par TUKEY, la boîte a pour hauteur la distance interquartile (Q3-Q1), et les moustaches sont basées généralement sur 1,5 fois la hauteur de la boîte. Dans ce cas, une valeur est atypique si elle dépasse de 1.5 fois l'écart interquartile au-dessous du 1er quartile ou au-dessus du 3ème quartile. En se basant sur les quartiles, c'est à dire des statistiques d'ordre, la médiane et l'écart interquartile ne sont jamais influencés par les valeurs extrêmes. La valeur 1.5 est, selon TUKEY, une valeur pragmatique (rule of thumb) qui a une raison probabiliste. Si une variable suit une distribution normale, alors la zone délimitée par la boîte et les moustaches devrait contenir 99,3 % des observations. On ne devrait donc trouver que 0.7% d'observations atypiques (outliers). Si le coefficient vaut 1, la probabilité serait de 0.957, et elle vaudrait 0.999 si le coefficient est égal à 2. Pour TUKEY la valeur 1.5 est donc un compromis pour retenir comme atypiques assez d'observations sans en retenir trop.

- Résidu(s) minimum et maximum : notés en vert (resp. rouge), ils représentent les résidus qui ont le plus petit (resp. plus grand) écart par rapport à 0. Cela nous permet aussi de voir pour chacune des observations lesquelles sont les mieux (resp. moins bien) prédites.

Un tutoriel sur la façon d'utiliser XGBOOST est disponible sur le Centre d'aide XLSTAT :

https://www.xlstat.com/demo/gbm_fr

Bibliographie

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794). New York, NY, USA: ACM. https://doi.org/10.1145/2939672.2939785

J. Friedman. Stochastic gradient boosting. Computational Statistics & Data Analysis, 38(4):367–378, 2002.

J.H. Friedman (2001). "Greedy function approximation: a gradient boosting machine." Annals of Statistics, pp. 1189–1232

S. Gey et J. M. Poggi, Boosting and instabillity for regression trees, Rap. tech. 36, Université de Paris Sud, Mathématiques, 2002.

Friedman J, Hastie T, Tibshirani R, et al. (2000). "Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors)." The annals of statistics, 28(2), 337–407.

Le Guen, M. (2001). La boîte à moustaches de TUKEY, un outil pour initier à la statistique. *Statistiquement votre-SFDS, (4)*, 1-3.

XGBoost Documentation: https://xgboost.readthedocs.io/en/stable/

T. Friedman, T. Hastie and R. Tibshirani ADDITIVE LOGISTIC REGRESSION: A STATISTICAL VIEW OF BOOSTING,The Annals of Statistics 2000, Vol. 28, No. 2, 337–407

J. H. Friedman, Stochastic gradient boosting, Computational Statisrics and Data Analysis 38 (2002).

Tests de corrélation/association

Tests de corrélation

Utilisez cet outil pour calculer les coefficients de corrélation de Pearson, Spearman, Kendall ou polychoriques entre au moins deux variables, et pour éventuellement déterminer si les corrélations sont significatives ou non. Des visualisations des matrices de corrélation sont aussi proposées.

Dans cette section :

Description

Quatre coefficients sont proposés pour calculer la corrélation entre des variables quantitatives continues, discrètes ou ordinales.

Le **coefficient de corrélation de Pearson** : cette statistique est le coefficient de corrélation le plus communément utilisé car bien adapté aux données quantitatives continues. Sa valeur est comprise entre -1 et 1, et il mesure le niveau de relation linéaire entre deux variables. Remarque : le coefficient de Pearson au carré, appelé R^2, donne une idée de la proportion de variabilité d'une variable explicable par l'autre. Les p-values calculées pour les coefficients de corrélation permettent de tester l'hypothèse nulle de corrélation non significativement différente de zéro entre les variables. Cependant, il convient d'être prudent car, si l'indépendance entre deux variables implique la nullité du coefficient de corrélation entre les variables, la réciproque n'est pas vraie : on peut avoir une corrélation proche de zéro entre deux variables parce que la relation n'est pas linéaire, ou parce qu'elle est complexe et nécessite la prise en compte d'autres variables.

Le **coefficient de corrélation de Spearman** (rho) : ce coefficient utilise les rangs des observations et non leur valeur en tant que telle. Comme pour le coefficient de Pearson, on peut aussi interpréter ce coefficient en termes de variabilité expliquée. Ici, il s'agit bien entendu de la variabilité des rangs.

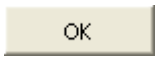
Le **coefficient de corrélation de Kendall** (tau) : comme pour le coefficient de Spearman, ce coefficient est aussi basé sur les rangs. Il est cependant conceptuellement très différent. Il peut être interprété comme en termes de probabilité : c'est la différence entre la probabilité pour que les variables varient dans le même sens et la probabilité pour qu'elles varient dans le sens

contraire. Lorsque le nombre d'observations est inférieur à 50 et qu'il n'y a pas d'ex-æquo, XLSTAT fournit la p-value exacte. Sinon une approximation est utilisée. Cette dernière est réputée fiable, dès lors qu'il y a plus de 8 observations.

Le **coefficient de corrélation polychorique** : ce coefficient permet de calculer la liaison entre deux variables qualitatives ordinales. L'hypothèse effectuée est que les variables ordinales sont obtenues grâce à la discrétisation de deux variables quantitatives non observées et suivant une loi normale. Le coefficient de corrélation polychorique mesure la liaison entre ces deux variables quantitatives non observées. Ce coefficient est souvent utilisé pour analyser les résultats d'une enquête où les réponses sont des variables qualitatives ordinales.

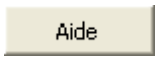
Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.



: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.

Onglet **Général**:

Tableau observations/variables : sélectionnez un tableau comprenant les observations. Si des en-têtes de colonne ont été sélectionnés pour les variables, veuillez vérifier que l'option « Libellés des variables » est activée. Pour les corrélations de Pearson, Spearman et Kendall, les variables doivent être quantitatives, pour les corrélations polychoriques les variables doivent être qualitatives ordinales.

Ordre des modalités (uniquement pour les corrélations polychoriques) : Vous pouvez sélectionner un tableau contenant l'ordre des modalités pour chaque variable. Ce tableau doit avoir autant de colonnes que le tableau contenant les variables. Si vous ne sélectionnez pas ce tableau, l'ordre des modalités pour chacune des variables est l'ordre lexicographique.

Poids : activez cette option si vous voulez pondérer les observations. Si vous n'activez pas cette option, les poids seront tous considérés comme valant 1. Les poids doivent être impérativement supérieurs ou égaux à 0. Si un en- tête de colonne a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

Type de corrélation : choisissez le type de corrélation à utiliser pour les calculs (voir la section description pour plus de détails).

Sous-échantillons : activez cette option pour sélectionner une colonne indiquant les noms ou les indices des sous-échantillons correspondant à chacune des observations. Les calculs des corrélations sont alors éffectués pour chacun des sous-échantillons.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des variables : activez cette option si la première ligne (ou colonne en mode lignes) des données sélectionnées (tableau observations/variables et poids) contient un libellé.

Niveau de signification (%) : entrez le niveau de signification qui permet de déterminer si les corrélations sont significatives ou non (valeur par défaut : 5%).

Onglet **Données manquantes** :

Ne pas accepter les valeurs manquantes : activez cette option pour que XLSTAT empêche la poursuite des calculs si des valeurs manquantes sont détectées.

Supprimer les observations : activez cette option pour supprimer les observations comportant des données manquantes.

Suppression par paire : activez cette option pour supprimer les observations comportant des données manquantes uniquement lorsque les variables impliquées dans les calculs comportent des données manquantes. Par exemple lors du calcul d'une corrélation entre deux variables, une observation ne sera ignorée que si la donnée correspondant à l'une des deux variables est manquante.

Estimer les données manquantes : activez cette option pour estimer les données manquantes avant le début des calculs.

- **Moyenne ou mode** : activez cette option pour estimer les données manquantes en utilisant la moyenne (variables quantitatives) ou le mode (variables qualitatives) pour les variables correspondantes.

- **Plus proche voisin** : activez cette option pour estimer les données manquantes d'une observation en recherchant le plus proche voisin de l'observation.

Onglet **Sorties**:

Statistiques descriptives : activez cette option pour afficher les statistiques descriptives pour les variables sélectionnées.

Corrélations : activez cette option pour afficher la matrice de corrélation correspondant au type de corrélation choisi dans l'onglet « Général ».

p-values : activez cette option pour calculer et afficher les p-values correspondant à chacune des corrélations. Si cette option est activée, les corrélations significatives au seuil de signification choisi seront affichées en gras dans la matrice des corrélations.

Coefficients de détermination (R2) : activez cette option pour calculer et afficher les coefficients de détermination qui sont les carrés des coefficients de corrélations.

Filtrer les variables avec R2 : cette option vous permet de filtrer l'affichage des variables, différentes options de filtrage sont disponibles :

- **R2 > (resp. R2 <)** : cette option permet de n'afficher que les variables ayant au moins un coefficient de détermination (R2) supérieur (resp. inférieur) au seuil que vous choisissez. Le seuil entré doit être compris entre 0 et 1.
- **p var avec Somme(R2) maximale (resp. minimale**) : cette option permet de n'afficher que les p variables (en précisant le p choisi) dont la somme des R2 avec toutes les autres variables est maximale (resp. minimale).

Trier les variables avec R2 : activez cette option si vous voulez trier les variables des matrices affichées en sortie de telle sorte que les variables fortement corrélées soient regroupées. Deux options de tri sont disponibles :

- **Méthode BEA** : la méthode BEA (Bond Energy Algorithm) développée initiallement par McCormick (1972) applique une permutation des lignes et des colonnes d'une matrice carrée afin que les variables présentant des corrélations similaires soient regroupées.
- **Méthode FPC** : la méthode FPC (First Principal Component) consiste à réaliser une ACP sur la matrice des coefficients de détermination R2. Une fois l'ACP réalisée, on réorganise les lignes et les colonnes de la matrice R2 de telle sorte que les variables soient triées dans l'ordre ascendant de leurs corrélations avec la première composante principale de l'ACP.

Onglet **Graphiques** :

Cartes des corrélations : plusieurs représentations d'une matrice des corrélations vous sont proposées.

- L'option « **Echelle bleu-rouge** » vous permet de représenter les corrélations faibles par des couleurs froides (bleu pour les corrélations proche de -1) et les corrélations élevées par des couleurs chaudes (rouge pour les corrélations proches de 1).

- L'option « **Noir et blanc** » vous permet soit de représenter en noir les corrélations positives et en blanc les corrélations négatives (la diagonale de 1 est représentée en gris), soit de représenter en noir les corrélations significativement non nulles, et en blanc les corrélations non significativement différentes de 0.

- L'option « **Motifs** » vous permet de représenter les corrélations positives par des traits montant de gauche à droite, et les corrélations négatives par des traits montant de droite à gauche. Plus la corrélation est élevée en valeur absolue, plus les traits sont espacés.

Nuages de points : activez cette option pour afficher les nuages de points pour toutes les combinaisons possibles de variables deux à deux.

- **Matrice de graphiques** : activez cette option pour afficher l'ensemble des combinaisons possibles de variables deux à deux sous la forme d'un tableau à deux entrées, avec en ligne et en colonne les différentes variables.

- **Histogrammes** : activez cette option pour que XLSTAT affiche les histogrammes des variables sur la diagonale de la matrice de graphiques.

- **Q-Q plots** : activez cette option pour que XLSTAT affiche les Q-Q plots des variables sur la diagonale de la matrice de graphiques.

- **Ellipses de confiance :** activez cette option pour afficher des ellipses de confiance. Les ellipses de confiance correspondent à un intervalle de confiance que vous pouvez spécifier pour une loi normale bivariée de même moyenne et de même matrice de covariance que les variables représentées en abscisse et en ordonnée.

Onglet **Image** :

Image : si vous avez choisi d'afficher les matrices de corrélation et/ou de détermination dans la feuille de résultats, vous pouvez choisir de les représenter sous la forme d'une image. Si une méthode de tri ou de filtrage a été sélectionnée dans l'onglet Sorties, les images des matrices prendront en compte uniquement les variables non filtrées et les afficheront sur l'image dans l'ordre du filtrage. Cette option peut être très utile lorsque vous disposez d'un grand nombre de variables afin de voir rapidement quelles variables présentent la même structure. Différentes options d'affichage des images sont disponibles :

- **Libellés des variables** : cette option permet d'afficher les libellés des variables au dessus de l'image.

- **Grille** : cette option permet l'affichage de traits fin entre les variables afin de les séparer visuellement sur l'image.

- **Légende** : cette option permet d'afficher une légende indiquant à quelles valeurs correspondent les couleurs sur l'image.

Résultats

La matrice de corrélation et le tableau des p-values sont affichés. Les cartes de corrélation permettent d'identifier d'éventuelles structures dans les corrélations, ou d'identifier rapidement les corrélations intéressantes.

Exemple

Un exemple de calcul du coefficient de corrélation de Spearman et du test de significativité correspondant est disponible sur le Centre d'aide XLSTAT à l'adresse suivante :

http://www.xlstat.com/demo-corrspf.htm

Bibliographie

Best D. J. and Roberts D. E. (1975). Algorithm AS 89: The upper tail probabilities of Spearman's rho. *Applied Statistics*, **24**, 377–379.

Best D.J. and Gipps P.G. (1974). Algorithm AS 71, Upper tail probabilities of Kendall's tau. *Applied Statistics*, **23**, 98-100.

Hollander M. and Wolfe D. A. (1973). Nonparametric Statistical Inference. John Wiley & Sons, New York.

Kendall M. (1955). Rank Correlation Methods, Second Edition. Charles Griffin and Company, London.

Lehmann E.L (1975). Nonparametrics: Statistical Methods Based on Ranks. Holden-Day, San Francisco.

McCormick and William T. (1972). Problem decomposition and data reorganization by a Clustering technique. *Operation Research*. **20(5)**, 993-1009.

Martinson E. O. and Hamdan M. A. (1975). Algorithm AS 87: Calculation of the polychoric estimate of correlation in contingency tables. *Journal of the Royal Statistical Society (Applied Statistics)*, **24(2)**, 272-278.

Coefficient RV

Utilisez cet outil pour calculer le coefficient RV entre deux tableaux de données quantitatives décrivant les mêmes observations.

Dans cette section :

Description

Cet outil permet de calculer le coefficient RV entre deux matrices de variables quantitatives décrivant les mêmes observations. Le coefficient RV est défini par (Robert and Escoufier, 1976; Schlich, 1996) :

$$RV(W_i, W_j) = \frac{trace(W_i, W_j)}{\sqrt{trace(W_i, W_i)trace(W_j, W_j)}}$$

où

$$trace(W_i, W_j) = \sum_{l,m} w_{lm}^i w_{lm}^j$$

est le coefficient de covariance généralisé entre les matrices W_i et W_j et

$$trace(W_i, W_i) = \sum_{l,m} [w_{lm}^i]^2$$

est la variance généralisée de la matrice W_i et $w_{l,m}^i$ est l'élément (l,m) de la matrice W_i.

Le coefficient RV est une généralisation du coefficient de Pearson élevé au carré. Le coefficient RV est compris entre 0 et 1. Plus le coefficient RV est proche de 1, plus les matrices W_i et W_j sont similaires.

XLSTAT propose :

- de calculer le coefficient RV entre deux matrices, en utilisant toutes les variables des deux matrices ;

- de choisir les k premières variables de chaque matrice et de calculer le coefficient RV entre les sous-matrices choisies.

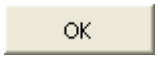
XLSTAT permet de tester si le coefficient RV obtenu est significativement différent de 0 ou non.

Deux méthodes de calcul de la p-value sont proposées par XLSTAT. L'utilisateur peut choisir entre un calcul basé sur une approximation Pearson III de la distribution de la statistique RV (Kazi-Aoual et *al.*, 1995), et un calcul basé sur des rééchantillonnages Monte-Carlo.

Remarque : la fonction XLSTAT_RVcoefficient permet de calculer le coefficient RV directement dans une feuille de calcul.

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.

Onglet **Général**:

Matrice A : Sélectionnez les données correspondant à N observations et P variables quantitatives. Si le libellé des colonnes a été sélectionné, vérifiez que l'option « Libellés des colonnes » est activée.

Matrice B Sélectionnez les données correspondant à N observations et Q variables quantitatives. Si le libellé des colonnes a été sélectionné pour la matrice A, il doit être sélectionné pour la matrice B.

Plage : Activez cette option pour afficher les résultats à partir d'une cellule située dans une feuille existante. Puis, sélectionnez la cellule.

Feuille : Activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : Activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des colonnes : Activez cette option si la première ligne des données sélectionnées (Tableau observations/variables, libellés des lignes) contient un libellé.

Libellés des lignes : Activez cette option si vous voulez utiliser des libellés d'observations. Si l'option « Libellés des colonnes » est activée, la première cellule de la sélection doit comprendre un en-tête.

Onglet **Options**:

Variables sélectionnées :

Toutes : Choisissez cette option pour calculer le coefficient RV entre les matrices A et B en utilisant toutes les variables des deux matrices.

Définies par l'utilisateur : Choisissez cette option pour calculer le coefficient RV entre des sous-matrices des matrices A et B ayant le même nombre de variables. Puis, entrez le nombre de variables à utiliser. Par exemple pour calculer le coefficient RV sur les deux premières variables (ou les deux premières dimensions dans le cas de résultats d'analyses multivariées), entrez 2 pour **de** et **à**. Pour calculer le coefficient RV pour une série de nombre de variables, entrez a pour **de** et b pour **à** où a<b. Par exemple pour calculer les coefficients RV pour les 2, 3 et 4 premières variables, entrez 2 pour **de** et 4 pour **à**.

Calcul des p-values :

Extrapolation : Choisissez cette option pour effectuer le calcul de la p-value associée au coefficient RV sur la base d'une approximation Pearson III de la distribution de la statistique RV.

Permutations : Choisissez cette option pour effectuer le calcul de la p-value sur la base de rééchantillonnages Monte-Carlo, et indiquez le nombre de permutations à effectuer ou le temps maximum de calcul.

Onglet **Données manquantes** :

Ne pas accepter les valeurs manquantes : activez cette option pour que XLSTAT empêche la poursuite des calculs si des valeurs manquantes sont détectées.

Supprimer les observations : activez cette option pour supprimer les observations comportant des données manquantes.

Onglet **Sorties**:

Coefficients RV : Activez cette option pour afficher le(s) coefficient(s) RV, le(s) coefficient(s) RV standardisé(s), et la(les) moyenne(s) et variance(s) de la distribution des coefficients RV.

Coefficients RV ajustés : Activer cette option pour afficher le(s) coefficients RV ajusté(s).

p-values : Activer cette option pour afficher la (les) p-value(s) associée(s) au coefficient(s) RV.

Onglet **Graphiques**:

Graphique des coefficients RV : Activer cette option pour afficher le graphique en barres des coefficients RV (avec un code couleur correspondant aux p-values associées si l'option **p-values** a été sélectionnée).

Résultats

Coefficients RV : dans ce tableau sont affichés le(s) coefficient(s) RV, le(s) coefficient(s) RV standardisé(s), et la(les) moyenne(s) et variance(s) de la distribution des coefficients RV, et afficher le(s) coefficients RV ajusté(s) la (les) p-value(s) associée(s) si demandé par l'utilisateur.

Diagramme en bâtons : Ce diagramme en bâtons permet de visualiser le(s) coefficient(s) RV (avec un code couleur correspondant aux p-values associées si demandé par l'utilisateur).

Exemple

Un exemple de calcul de coefficients RV est disponible sur le Centre d'aide XLSTAT :

http://www.xlstat.com/demo-rvf.htm

Bibliographie

Kazi-Aoual F., Hitier S., Sabatier R. and Lebreton J.-D. (1995). Refined approximations to permutations tests for multivariate inference. *Computational Statistics and Data Analysis*, **20**, 643–656.

Robert P. and Escoufier Y. (1976). A unifying tool for linear multivariate statistical methods: the RV-coefficient. *Applied Statistics*, **25**, 257–265.

Schlich P. (1996). Defining and validating assessor compromises about product distances and attribute correlations. In T, Næs, & E. Risvik (Eds.), Multivariate analysis of data in sensory sciences. New York: Elsevier.

Tests sur les tableaux de contingence (Khi², ...)

Utilisez cet outil pour étudier le degré d'association entre les lignes et les colonnes d'un tableau de contingence (tableau croisé), et pour tester l'indépendance entre les lignes et les colonnes.

Remarque : pour construire un tableau de contingence à partir de deux variables qualitatives, vous pouvez utiliser l'outil « Créer un tableau de contingence ».

Dans cette section :

Description

De nombreuses mesures d'association et plusieurs tests ont été proposés afin d'évaluer le lien entre les R lignes et les C colonnes d'un tableau de contingence.

Certaines mesures d'association ont été spécifiquement développées pour les tableaux 2×2. D'autres ont été mises au point pour le cas où les catégories des variables sont ordinales.

XLSTAT affiche systématiquement toutes les mesures. Néanmoins, les mesures concernant les variables ordinales ne pourront être interprétées que si les variables sont ordinales, et classées en ordre croissant dans le tableau de contingence.

Tests d'indépendance entre les lignes et les colonnes d'un tableau de contingence

- La statistique du **Khi² de Pearson** permet de tester l'indépendance entre les lignes et les colonnes du tableau en mesurant à quel point le tableau est éloigné (au sens du Khi²) de ce que l'on pourrait obtenir en moyenne, en conservant les mêmes sommes marginales. La statistique est donnée par :

$$\chi^2_P = \sum_{i=1}^{R}\sum_{j=1}^{C}\frac{(n_{ij}-f_{ij})^2}{f_{ij}}, \quad \text{avec } f_{ij} = \frac{n_{i.}n_{.j}}{n}, \quad n = \sum_{i=1}^{R}\sum_{j=1}^{C} n_{ij}, \quad n_{i.} = \sum_{j=1}^{C} n_{ij}, \quad n_{.j} = \sum_{i=1}^{R}$$

On montre que cette statistique suit une loi du Khi² à $(R-1)(C-1)$ degrés de liberté. Ce résultat étant asymptotique, il est prudent avant d'utiliser ce test de vérifier que :

- Que n est supérieur ou égal à 20,
- Qu'aucune somme marginale ($n_{i.}$ ou $n_{.j}$) n'est inférieure à 5

- Qu'au moins 80% des f_{ij} sont supérieurs à 5.

- Dans le cas où $R = 2$ et $C = 2$, une **correction de continuité** a été proposée par Yates (1934). On a alors :

$$\chi^2_Y = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{\left(|n_{ij} - f_{ij}| - 0.5\right)^2}{f_{ij}}$$

- Un test utilisant un **rapport de vraisemblances** a été proposé. Il utilise la statistique du G^2 de Wilks, et consiste à comparer la vraisemblance du tableau observé à celui d'un tableau moyen défini comme ci-dessus. On a :

$$G^2 = 2\sum_{i=1}^{R}\sum_{j=1}^{C} n_{ij}\log(n_{ij}/f_{ij})$$

Comme pour la statistique de Pearson, on montre que cette statistique suit une loi du Khi² à $(R-1)(C-1)$ degrés de liberté.

- Le **test exact de Fisher** permet de calculer la probabilité pour qu'un tableau montrant une association encore plus forte entre les lignes et les colonnes soit observé, les sommes marginales étant fixées, et sous hypothèse nulle d'indépendance entre les lignes et les colonnes. Dans le cas d'un tableau 2×2, l'indépendance est mesurée ici au travers du *odds ratio* (voir ci-dessous) qui est le rapport $\theta = (n_{11}n_{22})/(n_{12}n_{21})$. L'indépendance correspond au cas où $\theta = 1$. Il y a donc trois hypothèses alternatives possibles : l'hypothèse bilatérale $\theta \neq 1$, l'hypothèse unilatérale à gauche $\theta < 1$ et l'hypothèse unilatérale à droite $\theta > 1$

XLSTAT permet de calculer le test exact de Fisher bilatéral pour les tableaux $R \geq 2$ et $C \geq 2$. La méthode utilisée est celle de Mehta (1986) et Clarkson (1993). Elle peut échouer dans certains cas. L'utilisateur est alors prévenu.

- **Test Monte Carlo** : un test non paramétrique utilisant des simulations Monte Carlo permet de tester l'indépendance entre les lignes et les colonnes. Un nombre de simulations défini par l'utilisateur est effectué afin de générer les tableaux de contingence ayant les mêmes sommes marginales que le tableau observé. La statistique du Khi² de Pearson est calculée pour chacun des tableaux simulés. La p-value est alors déterminée en utilisant la distribution obtenue à partir des simulations.

Mesures d'association (1)

Une première série de coefficients d'association entre les lignes et les colonnes d'un tableau de contingence est proposée :

- Le coefficient **Phi de Pearson** permet de mesurer l'association entre les lignes et les colonnes d'un tableau $R \times C$. Dans le cas d'un tableau 2×2, sa valeur, comprise entre -1 et 1, est donnée par :

$$\phi_P = \frac{n_{11}n_{22} - n_{12}n_{21}}{\sqrt{n_{1.}n_{2.}n_{.1}n_{.2}}}$$

Lorsque $R > 2$ et/ou $C > 2$, il est compris entre 0 et le minimum des racines de $R - 1$ et $C - 1$. Il est alors donné par :

$$\phi_p = \sqrt{\chi^2_P / n}$$

- **Coefficient de contingence** : ce coefficient, aussi dérivé du Khi² de Pearson, est donné par :

$$C = \sqrt{\chi^2_P / (\chi^2_P + n)}$$

- Le coefficient **V de Cramer** est aussi dérivé du Khi² de Pearson. Dans le cas d'un tableau 2×2, sa valeur, comprise entre -1 et 1 est donnée par :

$$V = \phi_p$$

Lorsque $R > 2$ et/ou $C > 2$, il est compris entre 0 et 1 et sa valeur est alors donnée par :

$$V = \sqrt{\frac{\chi^2_P / n}{\min(R - 1, C - 1)}}$$

Plus V est proche de 0, plus les lignes et les colonnes sont indépendantes.

- **T de Tschuprow** : ce coefficient, aussi dérivé du Khi² de Pearson, est compris entre 0 et 1. Sa valeur est donnée par :

$$T = \sqrt{\frac{\chi^2_P / n}{(R - 1, C - 1)}}$$

Plus T est proche de 0, plus les lignes et les colonnes sont indépendantes.

- **Tau Goodman et Kruskal (L/C) et (C/L)** : ce coefficient, proche dans l'esprit du Khi² de Pearson, est asymétrique. Il permet de mesurer le degré de dépendance des lignes vis-à-vis des colonnes (L/C) ou vice versa (C/L).

- **Kappa de Cohen** : ce coefficient est utilisé pour les tableaux $R \times R$. Il est utile dans le cas où l'on veut étudier l'association entre deux échantillons appariés (par exemple, on pose la même question aux mêmes individus à deux instants différents). La valeur de Kappa est comprise entre 0 et 1 et vaut 1 lorsqu'il y a un lien total entre les deux variables (les réponses sont identiques aux deux instants).

- **Q de Yule** : ce coefficient est utilisé pour les tableaux 2×2. Il est calculé à partir des produits des données concordantes $(n_{11}n_{22})$ et des données discordantes $(n_{12}n_{21})$. Sa valeur est comprise en -1 et 1. Une valeur négative correspond à une discordance entre les deux variables, une valeur proche de 0 correspond à l'indépendance, et une valeur

proche de 1 à une concordance. Le Q de Yule est égal au Gamma de Goodman et Kruskal, lorsque ce dernier est calculé sur un tableau 2×2.

- **Y de Yule** : ce coefficient est utilisé pour les tableaux 2×2. Son calcul est similaire à celui du Q de Yule et sa valeur est aussi comprise entre -1 et 1.

Mesures d'association (2)

Une seconde série de coefficients d'association entre les lignes et les colonnes d'un tableau de contingence est proposée, avec le calcul d'intervalles de confiance autour des valeurs estimées. Ces intervalles de confiance font appel à des résultats asymptotiques. La fiabilité des intervalles dépend donc du nombre de données.

- **Gamma de Goodman et Kruskal** : ce coefficient permet de mesurer le degré de concordance entre deux variables ordinales, sur une échelle allant de -1 à 1.

- **Tau de Kendall** : ce coefficient, aussi appelé tau-b, permet de mesurer sur une échelle de -1 à 1 le degré de concordance entre deux variables ordinales. Contrairement au coefficient Gamma, le calcul du tau de Kendall permet de prendre en compte les ex æquo.

- **Tau de Stuart** : ce coefficient, aussi appelé tau-c, permet de mesurer sur une échelle de -1 à 1 le degré de concordance entre deux variables ordinales. Comme pour le tau de Kendall, le tau-c permet de prendre en compte les ex æquo. En outre, il permet d'effectuer un ajustement en fonction de la taille du tableau.

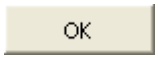
- **D de Somers (L/C) et (C/L)** : ce coefficient est une alternative asymétrique au tau de Kendall. Dans le cas (L/C) les lignes sont supposées dépendre des colonnes, et réciproquement dans le cas (C/L) ; la correction pour les ex æquo n'est apportée qu'à la variable « explicative ».

- **U de Theil (L/C) et (C/L)** : le coefficient asymétrique d'incertitude U de Theil (L/C) permet de mesurer quelle proportion de l'incertitude de la variable en ligne est expliquée par la variable en colonne, et réciproquement pour le cas C/L. Ce coefficient est compris entre 0 et 1. La version symétrique, aussi comprise entre 0 et 1 est calculée à partir des coefficients (L/C) et (C/L).

- **Odds ratio et Log(Odds ratio)** : le odds ratio est calculé dans le cas des tableaux 2×2 comme le rapport $\theta = (n_{11}n_{22})/(n_{12}n_{21})$. Odds signifie en anglais dans ce contexte « chance ». θ varie entre 0 et l'infini. θ peut être interprété comme le surcroît de chances d'être dans la colonne 1, lorsque l'on est dans la ligne 1 du tableau par rapport à ce que l'on aurait dans la ligne 2. Au cas $\theta = 1$ ne correspond aucun avantage. Lorsque $\theta > 1$, la probabilité sera θ fois supérieure pour la ligne 1 par rapport à la ligne 2. On calcule le logarithme du odds ratio parce que sa variance est aisément calculable, et parce que ce coefficient est symétrique autour de 0, ce qui permet d'obtenir un intervalle de confiance, d'où l'on déduit celui sur le odds ratio.

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.

Onglet **Général**:

Tableau de contingence : si le format de données choisi est « Tableau de contingence », sélectionnez un tableau croisé, avec les fréquences correspondant aux différentes catégories de deux variables qualitatives. Si les libellés des lignes et des colonnes du tableau ont été sélectionnés, veillez à ce que l'option « libellés inclus » soit activée.

Variable(s) ligne : si le format de données choisi est « Variables qualitatives », sélectionnez les données correspondant aux variables qualitatives qui seront les variables en ligne des tableaux de contingence créés. Si les libellés des variables ont été sélectionnés, veillez à ce que l'option « libellés des variables » soit bien activée.

Variable(s) colonne : si le format de données choisi est « Variables qualitatives », sélectionnez les données correspondant aux variables qualitatives qui seront les variables en colonne des tableaux de contingence créés. Si les libellés des variables ont été sélectionnés, veillez à ce que l'option « libellés des variables » soit bien activée.

Analyse par groupe : activez cette option puis sélectionnez ici les données d'appartenance à des groupes si vous souhaitez que les calculs soient effectués sur chaque groupe séparément.

Format des données : choisissez le format des données.

- **Tableau de contingence** : activez cette option si vos données sont disponibles sous la forme d'un tableau de contingence.
- **Variables qualitatives** : activez cette option si vos données se présentent sous la forme de deux variables qualitatives à partir desquelles sera généré un tableau de contingence.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés inclus : activez cette option si la première ligne et la première colonne des données sélectionnées contient un libellé.

Onglet **Options**:

Test du khi² : activez cette option pour effectuer le test du khi².

Test du rapport de vraisemblance : activez cette option pour effectuer le test du rapport de vraisemblance de Wilks.

Méthode Monte Carlo : activez cette option pour calculer la p-value en utilisant des simulations Monte Carlo.

Niveau de signification (%) : entrez le niveau de signification à utiliser pour les différents tests (valeur par défaut : 5%).

Test exact de Fisher : activez cette option pour calculer le test exact de Fisher. Dans le cas d'un tableau 2×2 vous pouvez choisir l'**hypothèse alternative**. Dans les autres cas, l'hypothèse alternative bilatérale sera automatiquement utilisée (voir la section description pour plus de détails).

Onglet **Données manquantes** :

Ne pas accepter les valeurs manquantes : activez cette option pour que XLSTAT empêche la poursuite des calculs si des valeurs manquantes sont détectées.

Remplacer les valeurs manquantes par 0 : activez cette option si vous considérez que les valeurs manquantes sont équivalentes à des 0.

Remplacer les valeurs manquantes par l'espérance : activez cette option si vous souhaitez remplacer les valeurs manquantes par leur espérance. L'espérance d'une valeur manquante est donnée par :

$$E(n_{ij}) = \frac{n_{i.} n_{.j}}{n}$$

où $n_{i.}$ est la somme sur les colonnes pour la ligne i, $n_{.j}$ est la somme sur les lignes pour la colonne j, et n est l'effectif total avant remplacement des valeurs manquantes.

Onglet **Sorties**:

Liste des combinaisons : activez cette option pour afficher la liste des différentes combinaisons possibles des deux variables qualitatives, ainsi que les effectifs correspondants.

Tableau de contingence : activez cette option pour afficher le tableau de contingence.

Inertie par case : activez cette option pour afficher les inerties correspondant à chacune des cellules du tableau de contingence.

Khi² par case : activez cette option pour afficher les Khi² correspondant à chacune des cellules du tableau de contingence.

Significativité par case : activez cette option pour afficher un tableau indiquant, pour chaque case, si la valeur observée est égale (=), inférieure (<) ou supérieure (>) à la valeur théorique, et pour effectuer un test (test exact de Fisher sur un tableau 2×2 ayant le même effectif total que le tableau complet, et les mêmes sommes marginales pour la case en question), afin de déterminer si l'écart à la valeur théorique est significatif ou non.

Coefficient d'association : activez cette option pour afficher les différents coefficients d'association calculés.

Effectifs observés : activez cette option pour afficher le tableau des effectifs observés. Ce tableau est presque identique au tableau de contingence, la différence venant des sommes marginales pour les lignes et les colonnes.

Effectifs théoriques : activez cette option pour afficher le tableau des effectifs théoriques estimés à partir des sommes marginales.

Proportions ou pourcentages / Ligne : activez cette option pour afficher le tableau des proportions ou pourcentages par ligne qui correspondent aux effectifs observés divisés par les sommes marginales des lignes.

Proportions ou pourcentages / Colonne : activez cette option pour afficher le tableau des proportions ou pourcentages par colonne qui correspondent aux effectifs observés divisés par les sommes marginales des colonnes.

Proportions ou pourcentages / Total : activez cette option pour afficher le tableau des proportions ou pourcentages calculés comme les effectifs observés divisés par l'effectif total.

Synthèse pour tous les groupes : activez cette option pour afficher un résumé de l'ensemble des tableaux de contingence

Onglet **Graphiques** :

Vue 3D du tableau de contingence / du tableau croisé : activez cette option pour afficher le diagramme en bâton en 3 dimensions correspondant au tableau de contingence ou au tableau croisé.

Tableau de contingence : activez cette option pour afficher le graphique associé au tableau de contingence.

Proportions ou pourcentages / Ligne : activez cette option pour afficher le graphique associé au tableau des proportions ou pourcentages par ligne.

Proportions ou pourcentages / Colonne : activez cette option pour afficher le graphique associé au tableau des proportions ou pourcentages par colonne.

Synthèse pour tous les groupes : activez cette option pour afficher les graphiques associés à chacun des groupes du tableau de synthèse.

Options des graphiques :

- **Type de graphique** :
  - **Groupé** : choisissez cette option pour afficher les graphiques sous forme de barres regroupées par modalité.
  - **Barres empilées** : choisissez cette option pour afficher les graphiques sous forme de barres empilées. Cela permet de comparer les effectifs ou les fréquences des sous-échantillons à ceux d'un échantillon complet.
- **Diagrammes en bâtons** :
  - **Effectifs** : choisissez cette option pour afficher l'effectif correspondant à chaque barre.
  - **Pourcentages** : choisissez cette option pour afficher le % de population correspondant à chaque barre.

Résultats

Les résultats calculés correspondent aux différentes statistiques et coefficients présentés dans la section description.

Exemple

Un exemple des tests sur les tableaux de contingence est disponible sur le Centre d'aide XLSTAT à l'adresse suivante :

http://www.xlstat.com/demo-ctof.htm

Bibliographie

Agresti A. (1990). Categorical data analysis. John Wiley & Sons, New York.

Agresti A. (1992). A survey of exact inference for contingency tables. *Statistical Science*, **7** (1), 131-177.

Everitt B. S. (1992). The Analysis of Contingency Tables, Second Edition. Chapman & Hall, New York.

Mehta C.R. and Patel N.R. (1986). Algorithm 643. FEXACT: A Fortran subroutine for Fisher's exact test on unordered r*c contingency tables. *ACM Transactions on Mathematical Software*, **12**, 154-161.

Clarkson D.B., Fan Y. and Joe H. (1993). A remark on algorithm 643: FEXACT: An algorithm for performing Fisher's exact test in r x c contingency tables. *ACM Transactions on Mathematical Software*, **19**, 484-488.

Fleiss J.L. (1981). Statistical Methods for Rates and Proportions, Second Edition. John Wiley & Sons, New York.

Saporta G. (1990). Probabilités, Analyse des Données et Statistique. Technip, Paris. 199-216.

Sokal R.R. and Rohlf F.J. (1995). Biometry. The Principles and Practice of Statistics in Biological Research, Third edition. Freeman, New York.

Theil H. (1972). Statistical Decomposition Analysis. North-Holland Publishing Company, Amsterdam.

Yates F. (1934). Contingency tables involving small numbers and the Chi- square test. *Journal of the Royal Statistical Society*, Suppl.1, 217-235.

Test de tendance de Cochran-Armitage

Utilisez cet outil pour tester si des proportions, éventuellement calculées à partir d'un tableau de contingence, peuvent être considérées comme variant linéairement en fonction d'une variable ordinale ou continue.

Dans cette section :

Description

Le test de Cochran-Armitage permet de tester si une série de proportions peut être considérée comme variant linéairement en fonction d'une variable ordinale ou continue.

Si X est la variable correspondand aux scores (les valeurs prises par la variable ordinale ou continue), la statistique permettant de tester la linéarité est donnée par :

$$z = \frac{\sum_{i=1}^{r} n_{i1}(X_i - \overline{X})}{\sqrt{p_{+1}(1-p_{+1})s^2}} \quad \text{avec} \quad s^2 = \sum_{i=1}^{r} n_{i+}(X_i - \overline{X})^2$$

Remarque : si X est une variable ordinale, la valeur du minimum de X n'a pas d'influence sur la valeur de z.

Dans le cas d'un test bilatéral, les hypothèses nulle (H_0) et alternative (H_a) sont les suivantes :

- $H_0 : z = 0$
- $H_a : z \neq 0$

Remarque : z est asymptotiquement distribuée comme une variable normale standard. Certains logiciels utilisent z^2 pour tester la linéarité. z^2 est alors distribuée suivant un χ^2 à 1 degré de liberté.

Dans le cas unilatéral, il faut distinguer le test unilatéral à gauche (ou inférieur) et le test unilatéral à droite (ou supérieur). Dans le test unilatéral à gauche, les hypothèses sont les suivantes :

- $H_0 : z = 0$
- $H_a : z < 0$

Si H_a est retenue on conclura que les proportions décroissent lorsque la variable score croît.

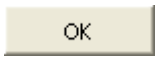
Pour le test unilatéral à droite, les hypothèses sont les suivantes :

- $H_0 : z = 0$
- $H_a : z > 0$

Si H_a est retenue on conclura que les proportions croissent lorsque la variable score croît.

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas (mode colonnes), XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite (mode lignes), XLSTAT considère que les variables sont en lignes et les observations en colonnes.

Onglet **Général** :

Tableau de contingence : sélectionnez un tableau de contingence. Si les libellés des colonnes du tableau ont été sélectionnés, veillez à ce que l'option « libellés des colonnes » soit activée.

Proportions : sélectionnez une colonne (ou une ligne dans le cas du mode lignes) contenant les proportions. Si un libellé de colonne a été sélectionné, veillez à ce que l'option « libellés des colonnes » soit activée.

Taille des échantillons : si vous avez sélectionné des proportions, vous devez ensuite sélectionner les effectifs correspondant. Si un libellé de colonne a été sélectionné, veillez à ce que l'option « libellés des colonnes » soit activée.

Libellés des lignes : activez cette option pour sélectionner les libellés des lignes.

Format des données :

- **Tableau de contingence** : activez cette option si vos données sont contenues dans un tableau de contingence.
- **Proportions** : activez cette option si vos données sont disponibles sous la forme de proportions et d'effectifs testés.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des colonnes : activez cette option si des en-têtes de colonne ont été sélectionnés dans les sélections.

Scores : vous pouvez choisir entre des scores ordinaux (1, 2, 3, ...) ou des scores dont vous entrez la valeur.

- **Ordinaux** : activez cette option pour utiliser des scores ordinaux.
- **Définis par l'utilisateur** : activez cette option pour sélectionnez les scores. Si un libellé de colonne a été sélectionné, veillez à ce que l'option « libellés des colonnes » soit activée.

Onglet **Options** :

Hypothèse alternative : choisissez l'hypothèse alternative à utiliser pour le test (voir description).

Niveau de signification (%) : entrez le niveau de signification à utiliser pour les différents tests (valeur par défaut : 5%).

p-value asymptotique : activez cette option pour calculer la p-value basée sur la distribution asymptotique de la statistique z.

Méthode Monte Carlo : activez cette option pour calculer la p-value en utilisant des simulations Monte Carlo. Entrez alors le nombre de simulations.

Onglet **Données manquantes** :

Ne pas accepter les valeurs manquantes : activez cette option pour que XLSTAT empêche la poursuite des calculs si des valeurs manquantes sont détectées.

Onglet **Sorties**:

Statistiques descriptives : activez cette option pour calculer et afficher les statistiques descriptives.

Onglet **Graphiques** :

Proportions : activez cette option pour afficher un graphique avec sur l'axe des abscisses les scores et sur l'axe des ordonnées les proportions.

Résultats

Les résultats affichés comprennent un tableau de synthèse reprenant l'ensemble des données, un graphique présentant les proportions en fonction des scores, puis les résultats et l'interprétation du test, basé sur la p-value calculée à partir de la distribution asymptotique, et la p-value calculée à partir de la distribution obtenue à partir des simulations Monte Carlo.

Exemple

Un exemple de calcul du test de Cochran-Armitage est disponible sur le Centre d'aide XLSTAT à l'adresse suivante :

http://www.xlstat.com/demo-cochran.htm

Bibliographie

Agresti A. (1990). Categorical Data Analysis. John Wiley and Sons, New York.

Armitage P. (1955). Tests for linear trends in proportions and frequencies. *Biometrics* ; **11**, 375-386.

Cochran W.G. (1954). Some methods for strengthening the common Chi-square tests, *Biometrics*, **10**, 417-451.

Snedecor G.W. and Cochran W.G. (1989). Statistical Methods, 8th Edition. Iowa State University Press, Ames.

Test de Mantel

Utilisez ce test pour calculer la corrélation linéaire entre deux matrices de proximité (test de Mantel simple), ou pour calculer la corrélation linéaire entre deux matrices connaissant leur corrélation avec une troisième matrice (test de Mantel partiel).

Dans cette section :

Description

Mantel (1967) a proposé une première statistique pour mesurer la corrélation entre deux matrices de proximité (similarité ou dissimilarité) symétriques A et B de taille n :

$$z(AB) = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} a_{ij} b_{ij}$$

La statistique standardisée de Mantel, plus pratique car variant entre -1 et 1, est le coefficient de corrélation de Pearson entre les deux matrices :

$$r(AB) = \frac{1}{n(n-1)/2 - 1} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \left(\frac{a_{ij} - \bar{a}}{s_a} \right) \left(\frac{b_{ij} - \bar{b}}{s_b} \right)$$

Remarques :

Dans le cas où les similarités ou les dissimilarités seraient de nature ordinale, on peut utiliser les coefficients de corrélation de Spearman ou de Kendall de manière identique.

Dans le cas où les matrices ne sont pas symétriques, le calcul est aussi possible.

S'il ne pose aucun problème de calculer un coefficient de corrélation entre des coefficients de proximité obtenus à partir de deux matrices de même taille, les tests habituellement utilisés à partir de ces coefficients ne peuvent pas être utilisés dans ce contexte, car ils nécessitent de pouvoir faire l'hypothèse d'indépendance entre les données, ce qui n'est pas le cas ici. Un test de permutation a donc été proposé pour permettre de déterminer si le coefficient de corrélation peut être considéré comme significativement différent de 0.

Dans le cas d'un test bilatéral, les hypothèses nulle (H0) et alternative (Ha) sont les suivantes :

- H0 : $r(AB) = 0$
- Ha : $r(AB) \neq 0$

Dans le cas unilatéral, il faut distinguer le test unilatéral à gauche (ou inférieur) et le test unilatéral à droite (ou supérieur). Dans le test unilatéral à gauche, les hypothèses sont les suivantes :

- H0 : $r(AB) = 0$
- Ha : $r(AB) < 0$

Pour le test unilatéral à droite, les hypothèses sont les suivantes :

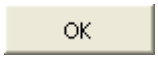
- H0 : $r(AB) = 0$
- Ha : $r(AB) > 0$

Le test de Mantel consiste à calculer quel coefficient de corrélation serait obtenu si l'on permutait les valeurs observées pour l'une des matrices. La p-value est alors déterminée à partir de la distribution des S coefficients r (AB) obtenus après S permutations. Dans le cas où n, le nombre de lignes et de colonnes des matrices, est inférieur à 10, toutes les permutations peuvent facilement être étudiées. Sinon, on est obligé de permuter la matrice de manière aléatoire, un grand nombre de fois, afin d'obtenir une distribution approchée.

Un test de Mantel pour plus de deux matrices a été proposé (Smouse *et al.*, 1986) : lorsque l'on dispose de trois matrices de proximité, A, B, C, la statistique partielle de Mantel pour les matrices A et B, connaissant C, est notée r(AB.C) et se calcule comme un coefficient de corrélation partiel. Afin de déterminer si le coefficient est significativement différent de 0 un test de Mantel partiel est calculé à partir de permutations.

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.

 : cliquez sur ce bouton pour rétablir les options par défaut.

 : cliquez sur ce bouton pour effacer les sélections de données.

 : cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.

Onglet **Général**:

Matrice A : sélectionnez la première matrice de proximité. Si les libellés des lignes et des colonnes du tableau ont été sélectionnés, veillez à ce que l'option « libellés inclus » soit activée.

Matrice B : sélectionnez la seconde matrice de proximité. Si les libellés des lignes et des colonnes du tableau ont été sélectionnés, veillez à ce que l'option « libellés inclus » soit activée.

Matrice C : activez cette option si vous voulez réaliser un test de Mantel partiel. Sélectionnez alors la troisième matrice de proximité. Si les libellés des lignes et des colonnes du tableau ont été sélectionnés, veillez à ce que l'option « libellés inclus » soit activée.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés inclus : activez cette option si la première ligne et la première colonne des données sélectionnées contient un libellé.

Onglet **Options**:

Hypothèse alternative : choisissez l'hypothèse alternative à utiliser pour le test (voir description).

Niveau de signification (%) : entrez la valeur du niveau de signification pour le test (valeur par défaut : 5%).

p-values exactes : activez cette option pour que XLSTAT tente dans la mesure du possible de calculer l'ensemble des permutations possibles pour obtenir une distribution exacte.

Nombre de permutations : entrez le nombre de permutations à réaliser dans le cas où toutes les permutations possibles ne pourraient être explorées.

Type de corrélation : choisissez le type de corrélation à utiliser pour le calcul de la statistique standardisée de Mantel.

Onglet **Données manquantes** :

Ne pas accepter les valeurs manquantes : activez cette option pour que XLSTAT empêche la poursuite des calculs si des valeurs manquantes sont détectées.

Onglet **Graphiques** :

Nuage de points : activez cette option pour afficher un graphique dont les points ont pour abscisse les valeurs de la matrice A et pour ordonnée les valeurs de la matrice B.

Histogramme : activez cette option pour afficher l'histogramme calculé pour la distribution de la statistique r(AB) à partir des permutations.

Résultats

Les résultats fournis correspondent à la statistique standardisée de Mantel la p-value correspondante pour l'hypothèse alternative choisie. Un début d'interprétation du test est affiché. L'histogramme de la distribution de r(AB) est affiché si l'option correspondante a été activée. La valeur observée pour r(AB) est indiquée sur l'histogramme.

Exemple

Un exemple de test de Mantel est disponible sur le Centre d'aide XLSTAT à l'adresse suivante :

http://www.xlstat.com/demo-mantelf.htm

Bibliographie

Legendre P. and Legendre L. (1998). Numerical Ecology, Second English Edition. Elsevier, Amsterdam.

Mantel N. (1967). A technique of disease clustering and a generalized regression approach. *Cancer Research*, **27**, 209-220.

Smouse P.E., Long J.C. and Sokal R.R. (1986). Multiple regression and correlation extension of the Mantel test of matrix correspondence. *Systematic Zoology*, **35**, 627-632.

Sokal R.R. and Rohlf F.J. (1995). Biometry. The Principles and Practice of Statistics in Biological Research. Third Edition. Freeman, New York.

Tests paramétriques

Tests t et z pour un échantillon

Utilisez cet outil pour comparer la moyenne d'un échantillon distribué suivant une loi normale à une valeur donnée.

Dans cette section :

Description

Soit un échantillon de moyenne $\hat{\mu}$. Pour comparer cette moyenne à une valeur de référence μ_0, deux tests paramétriques sont possibles :

- le test t de Student si on ne connaît pas la vraie variance de la population dont est extrait l'échantillon ; on utilise alors comme estimateur de la variance, la variance de l'échantillon s^2.
- le test z si on connaît la vraie variance σ^2 de la population.

Ces deux tests sont dits paramétriques car leur utilisation nécessite que l'on suppose que les échantillons sont distribués suivant une loi normale. Par ailleurs, on suppose aussi que les observations sont indépendantes et identiquement distribuées. La normalité de l'échantillon peut être préalablement testée grâce aux tests de normalité.

Trois types de tests sont possibles en fonction de l'hypothèse alternative choisie :

Pour le test bilatéral, les hypothèses nulle H0 et alternative Ha sont les suivantes :

- H0 : $\hat{\mu} = \mu_0$
- Ha : $\hat{\mu} \neq \mu_0$

Pour le test unilatéral à gauche, les hypothèses sont les suivantes :

- H0 : $\hat{\mu} = \mu_0$

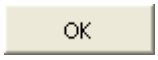
- Ha : $\hat{\mu} < \mu_0$

Pour le test unilatéral à droite, les hypothèses sont les suivantes :

- H0 : $\hat{\mu} = \mu_0$
- Ha : $\hat{\mu} > \mu_0$

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.

Onglet **Général** :

Données : sélectionnez les données sur la feuille Excel.

Format des données : choisissez le format des données.

- **Une colonne/ligne par échantillon** : activez cette option pour que XLSTAT considère que chaque colonne (mode colonnes) ou ligne (mode lignes) correspond à un échantillon. Vous pourrez ainsi en une seule fois tester l'hypothèse sur plusieurs échantillons.
- **Un échantillon** : activez cette option pour que XLSTAT considère que toutes les données sélectionnées, quelque soit le nombre de lignes ou de colonnes appartiennent au même échantillon.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des colonnes/lignes : activez cette option si la première ligne (mode colonnes) ou colonne (mode lignes) des données sélectionnées contient des libellés.

Test z : activez cette option pour utiliser un test z.

Test t de Student : activez cette option pour utiliser un test t de Student.

Onglet **Options**:

Hypothèse alternative : choisissez l'hypothèse alternative à utiliser pour le test (voir description).

Moyenne théorique : entrez la valeur de la moyenne théorique à laquelle la moyenne de l'échantillon doit être comparée.

Niveau de signification (%) : entrez la valeur du niveau de signification pour les tests (valeur par défaut : 5%).

Dans le cas où un test z est demandé, la valeur de variance de la population doit être entrée.

Variance pour le test z :

- **Estimée à partir de l'échantillon** : activez cette option pour que XLSTAT estime la variance de la population à partir des données de l'échantillon. Cela devrait en principe conduire à un test t, mais cette option est proposée à titre pédagogique.
- **Définie par l'utilisateur** : entrez la valeur de la variance connue de la population.

Onglet **Données manquantes** :

Ne pas accepter les valeurs manquantes : activez cette option pour que XLSTAT empêche la poursuite des calculs si des valeurs manquantes sont détectées.

Supprimer les observations : activez cette option pour supprimer les observations comportant des données manquantes.

Onglet **Sorties**:

Statistiques descriptives : activez cette option pour afficher les statistiques descriptives des échantillons.

Résultats

Les résultats affichés par XLSTAT correspondent aux différentes statistiques des tests sélectionnés, et à l'interprétation qui en découle.

Exemple

Bibliographie

Sincich T. (1996). Business Statistics by Example, 5th Edition. Prentice- Hall, Upper Saddle River.

Sokal R.R. and Rohlf F.J. (1995). Biometry. The Principles and Practice of Statistics in Biological Research. Third Edition. Freeman, New York.

Tomassone R., Dervin C. and Masson J.P. (1993). Biométrie. Modélisation de Phénomènes Biologiques. Masson, Paris.

Tests t et z pour deux échantillons

Utilisez cet outil pour comparer les moyennes de deux échantillons, indépendants ou appariés, distribués suivant une loi normale.

Dans cette section :

Description

L'utilisation des tests paramétriques t et z permet de comparer les moyennes de deux échantillons. La méthode de calcul est différente en fonction de la nature des échantillons. On distingue le cas où les échantillons sont indépendants (par exemple, dans le cas d'une comparaison du chiffre d'affaires annuel par magasin entre deux régions pour une chaîne de supermarchés), du cas où ils sont appariés (par exemple, dans le cas d'une comparaison, à l'intérieur d'une même région, des chiffres d'affaires annuels entre deux années).

Les tests t et z sont dits paramétriques car ils supposent que les échantillons sont distribués suivant des lois normales. Cette hypothèse pourra être testée à l'aide des tests de normalité.

Comparaison des moyennes de deux échantillons indépendants

Soit un échantillon E_1, comprenant n_1 observations, de moyenne $\hat{\mu_1}$ et de variance s_1^2. Soit un second échantillon E_2 indépendant de E_1, comprenant n_2 observations, de moyenne $\hat{\mu_2}$ et de variance s_2^2. Soit D la différence supposée entre les moyennes (D vaut 0 lorsque l'on suppose l'égalité).

Comme pour le cas des tests z et t sur un échantillon on utilise :

- le test t de Student si on ne connaît pas la vraie variance des populations dont sont extraits les échantillons ;
- le test z si on connaît la vraie variance σ^2 de la population.

Test t de Student

L'utilisation du test t de Student nécessite de décider préalablement si les variances des échantillons doivent être considérées comme étant égales ou non. XLSTAT propose d'utiliser un test F de Fisher afin de tester l'hypothèse d'égalité des variances, et de tenir compte du résultat du test pour la suite des calculs.

Si l'on considère que les deux échantillons ont la même variance, on estime la variance commune par :

$$s^2 = [(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2]/(n_1 + n_2 - 2)$$

La statistique du test est alors donnée par

$$t = \frac{(\hat{\mu_1} - \hat{\mu_2} - D)}{s\sqrt{1/n_1 + 1/n_2}}$$

La statistique t suit une loi de Student à $n_1 + n_2 - 2$ degrés de liberté.

Si l'on considère que les variances sont différentes la statistique est donnée par

$$t = \frac{(\hat{\mu_1} - \hat{\mu_2} - D)}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$$

Une modification du nombre de degrés de liberté a été proposée par Satterthwaite :

$$df = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}}$$

Remarque : lorsque $n_1 = n_2$, on a simplement $df = 2(n_1 - 1)$.

Cochran et Cox (1950) ont proposé une approximation pour déterminer la p-value. Elle est proposée en option dans XLSTAT.

Test z

Pour le test z, la variance s² de la population est supposée connue. L'utilisateur peut saisir cette valeur ou l'estimer à partir des données (ce dernier cas étant proposé uniquement à titre pédagogique). La statistique du test est donnée par :

$$z = \frac{(\hat{\mu_1} - \hat{\mu_2} - D)}{\sigma\sqrt{1/n_1 + 1/n_2}}$$

La statistique z suit une loi normale.

Comparaison des moyennes de deux échantillons appariés

Si deux échantillons sont appariés, ils sont nécessairement de même taille. Dans le cas où des valeurs seraient manquantes pour certaines observations, soit on supprime l'observation des deux échantillons, soit on estime les valeurs manquantes.

On étudie la moyenne des différences calculées pour les n observations. Si d est la moyenne des différences, s^2 la variance des différences, et D la différence supposée, la statistique du test t est donnée par :

$$t = \frac{(d - D)}{s/\sqrt{n}}$$

La statistique t suit une loi de Student à $n - 1$ degrés de liberté.

Pour le test z, la statistique, si σ^2 est la variance :

$$z = \frac{(d - D)}{\sigma/\sqrt{n}}$$

La statistique z suit une loi normale.

Hypothèses alternatives

Trois types de tests sont possibles en fonction de l'hypothèse alternative choisie :

Pour le test bilatéral, les hypothèses nulle H0 et alternative Ha sont les suivantes :

- H0 : $\hat{\mu_1} - \hat{\mu_2} = D$
- Ha : $\hat{\mu_1} - \hat{\mu_2} \neq D$

Pour le test unilatéral à gauche, les hypothèses sont les suivantes :

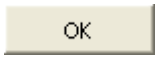
- H0 : $\hat{\mu_1} - \hat{\mu_2} = D$
- Ha : $\hat{\mu_1} - \hat{\mu_2} < D$

Pour le test unilatéral à droite, les hypothèses sont les suivantes :

- H0 : $\hat{\mu_1} - \hat{\mu_2} = D$
- Ha : $\hat{\mu_1} - \hat{\mu_2} > D$

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.

Onglet **Général** :

Données / Echantillon 1 : si le format de données sélectionné est « une colonne par variable », sélectionnez les données correspondant aux différents échantillons sur la feuille Excel. Si le format de données sélectionné est « une colonne par échantillon » ou « échantillons appariés », sélectionnez une colonne de données correspondant au premier échantillon.

Identifiant d'échantillon / Echantillon 2 : si le format de données sélectionné est « une colonne par variable», sélectionnez les données identifiant les échantillons auxquels les données sélectionnées correspondent (plusieurs colonnes peuvent être renseignées). Si le format de données sélectionné est « une colonne par échantillon » ou « échantillons appariés » sélectionnez une colonne de données correspondant au second échantillon.

Format des données : choisissez le format des données.

- **Une colonne/ligne par échantillon** : activez cette option pour sélectionner une colonne (ou ligne en mode lignes) par échantillon.
- **Une colonne/ligne par variable** : activez cette option pour que XLSTAT fasse autant de tests qu'il y a de colonnes/lignes, sachant que chaque colonne/ligne doit contenir le même nombre de lignes/colonnes, et qu'un identifiant d'échantillon permettant d'affecter chaque observation à un échantillon doit par ailleurs être sélectionné.
- **Echantillons appariés** : activez cette option pour faire des tests sur échantillons appariés. Vous devez alors sélectionner une colonne (ou ligne en mode lignes) par échantillon, tout en veillant à ce que les échantillons soient de même taille.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des colonnes/lignes : activez cette option si la première ligne (mode colonnes) ou colonne (mode lignes) des données sélectionnées contient des libellés.

Test z : activez cette option pour utiliser un test z.

Test t de Student : activez cette option pour utiliser un test t de Student.

Onglet **Options** :

Hypothèse alternative : choisissez l'hypothèse alternative à utiliser pour le test (voir description).

Différence supposée (D) : entrez la valeur de la différence supposée entre les moyennes des échantillons.

Niveau de signification (%) : entrez la valeur du niveau de signification pour les tests (valeur par défaut : 5%).

Deux méthodes de calcul de la p-value sont proposées. Choisissez la méthode **asymptotique** qui est la plus classiquement utilisée, ou la méthode **Monte Carlo**. Dans le cas de la méthode Monte Carlo, vous pouvez préciser quel temps maximum vous souhaitez consacrer au calcul de la p-value.

Poids : activez cette option si vous voulez pondérer les observations. Cette option n'est visible que si vous avez choisi le format de données « une colonne par variable » ou un test apparié. Si vous n'activez pas cette option, les poids seront tous considérés comme valant 1. XLSTAT prend en compte ces poids pour les calculs des degrés de libertés. Les poids doivent être impérativement supérieurs ou égaux à 0. Si un en-tête de colonne a été sélectionné, veuillez vérifier que l'option « Libellés des colonnes/lignes » est activée.

Dans le cas où un test z est demandé, la valeur de variance connue des populations, ou, dans le cas d'un test sur échantillons appariés, la variance de la différence, doit être entrée.

Variances pour le test z :

- **Estimées à partir des échantillons** : activez cette option pour que XLSTAT estime la variance de la population à partir des données des échantillons. Cela devrait en principe conduire à un test t, mais cette option est proposée à titre pédagogique.
- **Définies par l'utilisateur** : entrez la valeur des variances connues des populations.

Variances des échantillons pour le test t :

- **Supposer l'égalité** : activez cette option pour considérer que la variance des échantillons est égale.
- **Cochran-Cox** : activez cette option pour calculer la p-value en utilisant la méthode de Cochran et Cox dans le cas où les variances ne sont pas supposées égales.
- **Utiliser un test F** : activez cette option pour utiliser le test F de Fisher afin de déterminer si les variances des deux échantillons peuvent être considérées comme étant égales ou non.

Echantillons multiples : Dans le cas où l'option *Une colonne/ligne par variable* est sélectionnée, et que plusieurs colonnes ont été renseignées dans le champs *Identifiants d'échantillon* deux choix sont proposés :

- **Fusionner les échantillons** : activez cette option pour fusionner les différentes colonnes renseignées dans le champs *Identifiants d'échantillon*. Les tests seront réalisés suivant ce nouveau vecteur.
- **Traiter indépendamment** : activez cette option pour répliquer l'analyse de manière indépendante pour chacune des colonnes renseignées dans le champs *identifiants d'échantillon*.

Onglet **Données manquantes** :

Ne pas accepter les valeurs manquantes : activez cette option pour que XLSTAT empêche la poursuite des calculs si des valeurs manquantes sont détectées.

Ignorer les données manquantes : activez cette option pour ignorer les données manquantes. Cette option n'est disponible que si le format *une colonne par échantillon* a été sélectionnée.

Supprimer les observations :

- **Pour l'échantillon correspondant** : activez cette option pour supprimer les observations contenant des données manquantes, uniquement pour l'échantillon correspondant.
- **Pour tous les échantillons** : activez cette option pour supprimer les observations comportant des données manquantes, pour l'ensemble des échantillons.

Onglet **Sorties** :

Statistiques descriptives : activez cette option pour afficher les statistiques descriptives des échantillons.

Intervalle de confiance : activez cette option pour afficher l'intervalle de confiance autour de la différence des moyennes.

Résultats détaillés : si le test est calculé pour plusieurs variables, activez cette option pour afficher les résultats détaillés pour chacun des tests.

Synthèse des comparaisons : si le test est calculé pour plusieurs variables, activez cette option pour afficher la synthèse des comparaisons pour les différentes variables.

Dans le cas où nous avons plusieurs identifiants d'echantillons et qu'ils sont tous binaires (seulement deux groupes), la synthèse comprend également le graphique des p-valeurs associé.

Onglet **Graphiques** :

Diagramme de dominance : activez cette option pour afficher un diagramme de dominance afin de comparer visuellement les échantillons.

Distributions : activez cette option pour afficher la distribution de la statistique utilisée pour prendre la décision.

Résultats

Les résultats affichés par XLSTAT correspondent aux différentes statistiques des tests sélectionnés, et à l'interprétation qui en découle.

Le diagramme de dominance permet de comparer visuellement deux échantillons. Le premier échantillon est représenté sur l'axe des abscisses et le second sur l'axe des ordonnées. Pour construire ce diagramme, les données des échantillons sont d'abord triées. Lorsqu'une observation du second échantillon est supérieure à une observation du premier échantillon, un « + » est affiché. Lorsqu'une observation du second échantillon est inférieure à une observation du premier échantillon, un « - » est affiché. Dans le cas d'un ex aequo, un « o » est affiché.

Exemple

Un exemple de test de Student pour comparer les moyennes de deux échantillons est disponible sur

http://www.xlstat.com/demo-ttestf.htm

Bibliographie

Cochran W.G. and Cox G.M. (1950). Experimental Designs. John Wiley and Sons, New York.

Satterthwaite F.W. (1946). An approximate distribution of estimates of variance components. *Biometrics Bulletin*, **2**, 110 -114.

Sincich T. (1996). Business Statistics by Example, 5th Edition. Prentice- Hall, Upper Saddle River.

Sokal R.R. and Rohlf F.J. (1995). Biometry. The Principles and Practice of Statistics in Biological Research. Third Edition. Freeman, New York.

Tomassone R., Dervin C. and Masson J.P. (1993). Biométrie. Modélisation de Phénomènes Biologiques. Masson, Paris.

Tests de comparaison de moyennes pour k échantillons

Si vous souhaitez comparer les moyennes de k échantillons, vous devez utiliser l'outil d'ANOVA qui permet d'utiliser les tests de comparaisons multiples.

Test de la variance pour un échantillon

Utilisez cet outil pour comparer la variance d'un échantillon distribué suivant une loi normale à une valeur donnée.

Dans cette section :

Description

Soit un échantillon de n observations indépendantes distribuées suivant une loi normale. On montre alors que variance de l'échantillon s² suit, à un facteur de proportionnalité près, une loi du χ^2 à $n-1$ degrés de liberté.

$$s^2 \sim \frac{\sigma^2}{n-1}\chi^2_{n-1}$$

où σ^2 est la variance théorique de l'échantillon. Cette relation permet d'obtenir un intervalle de confiance autour de la valeur estimée de la variance.

Pour comparer la variance observée à une valeur de référence σ_0^2, un test paramétrique est proposé. Il est basé sur le calcul de la statistique

$$\chi_0^2 = (n-1)\frac{s^2}{\sigma_0^2}$$

qui suit une loi du χ^2 à $n-1$ degrés de liberté.

Ce test est dit paramétrique car son utilisation nécessite que l'on suppose que l'échantillon est distribué suivant une loi normale. Par ailleurs, on suppose aussi que les observations sont indépendantes et identiquement distribuées. La normalité de l'échantillon peut être préalablement testée grâce aux tests de normalité.

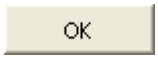
Trois types de tests sont possibles en fonction de l'hypothèse alternative choisie :

- Pour le test bilatéral, les hypothèses nulle H0 et alternative Ha sont les suivantes :
- H0 : $s^2 = \sigma_0^2$
- Ha : $s^2 \neq \sigma_0^2$

- Pour le test unilatéral à gauche, les hypothèses sont les suivantes :
- H0 : $s^2 = \sigma_0^2$
- Ha : $s^2 < \sigma_0^2$
- Pour le test unilatéral à droite, les hypothèses sont les suivantes :
- H0 : $s^2 = \sigma_0^2$
- Ha : $s^2 > \sigma_0^2$

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.

Onglet **Général**:

Données : sélectionnez les données sur la feuille Excel.

Format des données : choisissez le format des données.

- **Une colonne/ligne par échantillon** : activez cette option pour que XLSTAT considère que chaque colonne (mode colonnes) ou ligne (mode lignes) correspond à un échantillon. Vous pourrez ainsi en une seule fois tester l'hypothèse sur plusieurs échantillons.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des colonnes/lignes : activez cette option si la première ligne (mode colonnes) ou colonne (mode lignes) des données sélectionnées contient des libellés.

Onglet **Options**:

Hypothèse alternative : choisissez l'hypothèse alternative à utiliser pour le test (voir description).

Variance théorique : entrez la valeur de la variance théorique à laquelle la variance de l'échantillon doit être comparée.

Niveau de signification (%) : entrez la valeur du niveau de signification pour les tests (valeur par défaut : 5%).

Onglet **Données manquantes** :

Ne pas accepter les valeurs manquantes : activez cette option pour que XLSTAT empêche la poursuite des calculs si des valeurs manquantes sont détectées.

Supprimer les observations : activez cette option pour supprimer les observations comportant des données manquantes.

Onglet **Sorties**:

Statistiques descriptives : activez cette option pour afficher les statistiques descriptives des échantillons.

Résultats

Les résultats affichés par XLSTAT correspondent à l'intervalle de confiance calculé autour de la variance de l'échantillon et au test de comparaison de la variance observée à celle de l'échantillon.

Exemple

Un exemple de test de la variance pour un échantillon est disponible sur le Centre d'aide XLSTAT à l'adresse suivante :

http://www.xlstat.com/demo-variancef.htm

Bibliographie

Cochran W. G. (1934). The distribution of quadratic forms in a normal system, with applications to the analysis of covariance. *Mathematical Proceedings of the Cambridge Philosophical Society*, **30(2)**, 178-191.

Montgomery D. C. and Runger G. C. (2002). Applied Statistics and Probability for Engineers (3rd edition). John Wiley & Sons, Inc.

Comparaison des variances de deux échantillons

Utilisez cet outil pour comparer les variances de deux échantillons.

Dans cette section :

Description

Trois tests paramétriques sont proposés pour la comparaison des variances de deux échantillons. Soit un échantillon E_1, comprenant n_1 observations, de variance s_1^2. Soit un second échantillon E_2, comprenant n_2 observations, de variance s_2^2. XLSTAT propose trois tests pour comparer les variances des deux échantillons.

Test F de Fisher

Soit R le rapport supposé entre les variances (R vaut 1 lorsque l'onsuppose l'égalité).

La statistique F du test est donnée par :

$$F = \frac{s_1^2}{Rs_2^2}$$

Cette statistique suit une loi de Fisher à $(n_1 - 1)$ et $(n_2 - 1)$ de degrés de liberté si les deux échantillons suivent une loi normale.

Trois types de tests sont possibles en fonction de l'hypothèse alternative choisie :

Pour le test bilatéral, les hypothèses nulle H0 et alternative Ha sont les suivantes :

- H0: $s_1^2 = s_2^2 R$
- Ha: $s_1^2 \neq s_2^2 R$

Pour le test unilatéral à gauche, les hypothèses sont les suivantes :

- H0: $s_1^2 = s_2^2 R$
- Ha: $s_1^2 < s_2^2 R$

Pour le test unilatéral à droite, les hypothèses sont les suivantes :

- H0: $s_1^2 = s_2^2 R$
- Ha: $s_1^2 > s_2^2 R$

Test de Levene

Le test de Levene peut être utilisé pour comparer deux variances ou plus. C'est un test bilatéral pour lequel les hypothèses nulle et alternative sont dans le cas où deux variances sont comparées :

- H0: $s_1^2 = s_2^2 R$
- Ha: $s_1^2 \neq s_2^2 R$

La statistique de ce test est plus complexe que celle du test de Fisher et fait intervenir les écarts absolus à la moyenne (article original de Levene, 1960) ou à la médiane (Brown et Forsythe, 1974). L'utilisation de la moyenne est recommandée pour les distributions symétriques, à queues moyennement épaisses. L'utilisation de la médiane est recommandée pour les distributions asymétriques.

La statistique de Levene suit une loi de Fisher à 1 et $n_1 + n_2 - 2$ degrés de liberté.

Test d'homogénéité des variances de Bartlett

Le test de Bartlett peut être utilisé pour comparer deux variances ou plus. Ce test est sensible à la normalité des données. Autrement dit, si l'hypothèse de normalité des données semble fragile, on utilisera plutôt le test de Levene ou de Fisher. En revanche, le test de Bartlett est plus performant si les échantillons suivent une loi normale.

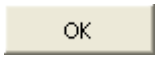
Il s'agit aussi d'un test bilatéral qui peut être utilisé avec deux variances ou plus. Dans le cas où deux variances sont comparées les hypothèses sont :

- H0: $s_1^2 = s_2^2 R$
- Ha: $s_1^2 \neq s_2^2 R$

La statistique de Bartlett suit une loi du χ^2 à 1 degré de liberté.

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

Annuler : cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

Aide : cliquez sur ce bouton pour afficher l'aide.

 : cliquez sur ce bouton pour rétablir les options par défaut.

 : cliquez sur ce bouton pour effacer les sélections de données.

 : cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.

Onglet **Général**:

Données / Echantillon 1 : si le format de données sélectionné est « une colonne par variable », sélectionnez les données correspondant aux différents échantillons sur la feuille Excel. Si le format de données sélectionné est « une colonne par échantillon », sélectionnez une colonne de données correspondant au premier échantillon.

Identifiant d'échantillon / Echantillon 2 : si le format de données sélectionné est « une colonne par variable», sélectionnez les données identifiant les deux échantillons auxquels les données sélectionnées correspondent. Si le format de données sélectionné est « une colonne par échantillon », sélectionnez une colonne de données correspondant au second échantillon.

Format des données : choisissez le format des données.

- **Une colonne/ligne par échantillon** : activez cette option pour sélectionner une colonne (ou ligne en mode lignes) par échantillon.
- **Une colonne/ligne par variable** : activez cette option pour que XLSTAT fasse autant de tests qu'il y a de colonnes/lignes, sachant que chaque colonne/ligne doit contenir le même nombre de lignes/colonnes, et qu'un identifiant d'échantillon permettant d'affecter chaque observation à un échantillon doit par ailleurs être sélectionné.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des colonnes/lignes : activez cette option si la première ligne (mode colonnes) ou colonne (mode lignes) des données sélectionnées contient des libellés.

Test F de Fisher : activez cette option pour utiliser un test F de Fisher (voir description).

Test de Levene : activez cette option pour utiliser le test de Levene (voir description).

- **Moyenne** : activez cette option pour utiliser le test de Levene basé sur la moyenne.
- **Médiane** : activez cette option pour utiliser le test de Levene basé sur la médiane.

Test Bartlett : activez cette option pour utiliser le test de Bartlett (voir description).

Onglet **Options**:

Hypothèse alternative : choisissez l'hypothèse alternative à utiliser pour le test (voir description).

Rapport supposé (R) : entrez la valeur du rapport supposé entre les variances des échantillons.

Niveau de signification (%) : entrez la valeur du niveau de signification pour les tests (valeur par défaut : 5%).

Deux méthodes de calcul de la p-value sont proposées. Choisissez la méthode **asymptotique** ou **Monte Carlo**. Dans le cas de la méthode Monte Carlo, vous pouvez préciser quel temps maximum vous souhaitez consacrer au calcul de la p-value.

Onglet **Données manquantes** :

Ne pas accepter les valeurs manquantes : activez cette option pour que XLSTAT empêche la poursuite des calculs si des valeurs manquantes sont détectées.

Supprimer les observations : activez cette option pour supprimer les observations comportant des données manquantes.

Onglet **Sorties**:

Statistiques descriptives : activez cette option pour afficher les statistiques descriptives des échantillons.

Intervalle de confiance : activez cette option pour afficher l'intervalle de confiance autour de la statistique.

Résultats détaillés : si le test est calculé pour plusieurs variables, activez cette option pour afficher les résultats détaillés pour chacun des tests.

Synthèse des comparaisons : si le test est calculé pour plusieurs variables, activez cette option pour afficher la synthèse des comparaisons pour les différentes variables.

Onglet **Graphiques** :

Distributions : activez cette option pour afficher la distribution de la statistique utilisée pour prendre la décision.

Résultats

Les résultats affichés par XLSTAT correspondent aux différentes statistiques des tests sélectionnés, et à l'interprétation qui en découle.

Exemple

Un exemple de test de Fisher pour comparer les variances de deux échantillons est disponible sur

http://www.xlstat.com/demo-ftestf.htm

Bibliographie

Brown M. B. and Forsythe A. B. (1974). Robust tests for the equality of variances. *Journal of the American Statistical Association*, **69**, 364-367.

Levene H. (1960). In Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling, I. Olkin et al. Editors. Stanford University Press, 278-292.

Sokal R.R. and Rohlf F.J. (1995). Biometry. The Principles and Practice of Statistics in Biological Research. Third Edition. Freeman, New York.

Tomassone R., Dervin C. and Masson J.P. (1993). Biométrie. Modélisation de Phénomènes Biologiques. Masson, Paris.

Comparaison des variances de k échantillons

Utilisez cet outil pour comparer les variances de k échantillons.

Dans cette section :

Description

Deux tests paramétriques sont proposés pour la comparaison des variances de k échantillons ($k = 2$). Soit k échantillons $E_1, E_2, \ldots, E_k$, comprenant $n_1, n_2, \ldots, n_k$ observations et de variance $s_1^2 = s_2^2 = \ldots = s_k^2$.

Test de Levene

Le test de Levene peut être utilisé pour comparer deux variances ou plus. C'est un test bilatéral pour lequel les hypothèses nulle et alternative sont :

- H0 : $s_1^2 = s_2^2 = \ldots = s_k^2$
- Ha : il existe au moins un couple (i, j) tel que $s_i^2 \neq s_j^2$

La statistique de ce test fait intervenir les écarts absolus à la moyenne (article original de Levene, 1960) ou à la médiane (Brown et Forsythe, 1974). L'utilisation de la moyenne est recommandée pour les distributions symétriques, à queues moyennement épaisses. L'utilisation de la médiane est recommandée pour les distributions asymétriques.

La statistique de Levene suit une loi de Fisher à $k - 1$ et $n1 + n2 - 2$ degrés de liberté.

Test d'homogénéité des variances de Bartlett

Le test de Bartlett peut être utilisé pour comparer deux variances ou plus. Ce test est sensible à la normalité des données. Autrement dit, si l'hypothèse de normalité des données semble fragile, on utilisera plutôt le test de Levene ou de Fisher. En, revanche le test de Bartlett est plus performant si les échantillons suivent une loi normale.

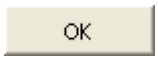
Il s'agit aussi d'un test bilatéral qui peut être utilisé avec deux variances ou plus. Dans le cas où deux variances sont comparées les hypothèses sont :

- H0 : $s_1^2 = s_2^2 = \ldots = s_k^2$
- Ha : il existe au moins un couple (i, j) tel que $s_i^2 \neq s_j^2$

La statistique de Bartlett suit une loi du Khi² à $k - 1$ degrés de liberté.

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.

Onglet **Général**:

Données / Echantillon 1 : si le format de données sélectionné est « une colonne par variable », sélectionnez les données correspondant aux différents échantillons sur la feuille Excel. Si le format de données sélectionné est « une colonne par échantillon », sélectionnez une colonne de données correspondant au premier échantillon.

Identifiant d'échantillon / Echantillon 2 : si le format de données sélectionné est « une colonne par variable», sélectionnez les données identifiant les k échantillons auxquels les données sélectionnées correspondent. Si le format de données sélectionné est « une colonne par échantillon » sélectionnez une colonne de données correspondant au second échantillon.

Format des données : choisissez le format des données.

- **Une colonne/ligne par échantillon** : activez cette option pour sélectionner une colonne (ou ligne en mode lignes) par échantillon.
- **Une colonne/ligne par variable** : activez cette option pour que XLSTAT fasse autant de tests qu'il y a de colonnes/lignes, sachant que chaque colonne/ligne doit contenir le même nombre de lignes/colonnes, et qu'un identifiant d'échantillon permettant d'affecter chaque observation à un échantillon doit par ailleurs être sélectionné.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des colonnes/lignes : activez cette option si la première ligne (mode colonnes) ou colonne (mode lignes) des données sélectionnées contient des libellés.

Test de Levene : activez cette option pour utiliser le test de Levene (voir description).

- **Moyenne** : activez cette option pour utiliser le test de Levene basé sur la moyenne.
- **Médiane** : activez cette option pour utiliser le test de Levene basé sur la médiane.

Test Bartlett : activez cette option pour utiliser le test de Bartlett (voir description).

Onglet **Options**:

Niveau de signification (%) : entrez la valeur du niveau de signification pour les tests (valeur par défaut : 5%).

Deux méthodes de calcul de la p-value sont proposées. Choisissez la méthode **asymptotique** ou **Monte Carlo**. Dans le cas de la méthode Monte Carlo, vous pouvez préciser quel temps maximum vous souhaitez consacrer au calcul de la p-value.

Onglet **Données manquantes** :

Ne pas accepter les valeurs manquantes : activez cette option pour que XLSTAT empêche la poursuite des calculs si des valeurs manquantes sont détectées.

Supprimer les observations : activez cette option pour supprimer les observations comportant des données manquantes.

Onglet **Sorties**:

Statistiques descriptives : activez cette option pour afficher les statistiques descriptives des échantillons.

Résultats détaillés : si le test est calculé pour plusieurs variables, activez cette option pour afficher les résultats détaillés pour chacun des tests.

Synthèse des comparaisons : si le test est calculé pour plusieurs variables, activez cette option pour afficher la synthèse des comparaisons pour les différentes variables.

Onglet **Graphiques** :

Distributions : activez cette option pour afficher la distribution de la statistique utilisée pour prendre la décision.

Résultats

Les résultats affichés par XLSTAT correspondent aux différentes statistiques des tests sélectionnés, et à l'interprétation qui en découle.

Exemple

Bibliographie

Brown M. B. and Forsythe A. B. (1974). Robust tests for the equality of variances. *Journal of the American Statistical Association*, **69**, 364-367.

Levene H. (1960). In Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling, I. Olkin et al. Editors. Stanford University Press, 278-292.

Sokal R.R. and Rohlf F.J. (1995). Biometry. The Principles and Practice of Statistics in Biological Research. Third Edition. Freeman, New York.

Tomassone R., Dervin C. and Masson J.P. (1993). Biométrie. Modélisation de Phénomènes Biologiques. Masson, Paris.

Tests multidimensionnels (Mahalanobis, ...)

Utilisez cet outil pour comparer deux groupes, ou plus, décrits par plusieurs variables quantitatives.

Dans cette section :

Description

Les tests mis en œuvre dans cet outil permettent de comparer des échantillons décrits par plusieurs variables. Par exemple, au lieu de comparer les moyennes de deux échantillons comme avec le test de Student, on compare ici simultanément pour les mêmes échantillons plusieurs moyennes mesurées pour différentes variables.

Par rapport à une procédure qui consisterait à faire autant de tests de Student qu'il y a de variables, la méthode proposée ici présente l'avantage d'utiliser la structure de covariance entre les variables et d'obtenir une conclusion globale. Il se peut que deux échantillons diffèrent pour une variable avec un test de Student, mais qu'au global on ne puisse rejeter l'idée qu'ils sont finalement semblables.

Distance de Mahalanobis

La distance de Mahalanobis, du nom du statisticien Indien Prasanta Chandra Mahalanobis (1893-1972), permet de calculer la distance entre deux points dans un espace à p dimensions, en tenant compte de la structure de variance- covariance sur ces p dimensions. Le carré de cette distance est défini par :

$$d_M^2 = (\vec{x}_1 - \vec{x}_2)' \Sigma^{-1} (\vec{x}_1 - \vec{x}_2)$$

Autrement dit, c'est le produit de la transposée du vecteur des différences de coordonnées pour les p dimensions entre les deux points, multiplié par l'inverse de la matrice de covariance, multiplié par le vecteur des différences. La distance euclidienne, correspond à la distance de Mahalanobis dans le cas où la matrice de covariance est la matrice identité, c'est-à-dire que les variables sont centrées réduites et indépendantes.

La distance de Mahalanobis peut être utilisée pour comparer deux groupes, car la statistique du T^2 de Hotelling définie par :

$$T^2 = \frac{n_1 n_2}{n_1 + n_2} d_M^2$$

suit une loi de Hotelling, à condition que les échantillons suivent une loi normale pour les différentes variables. La statistique F utilisée pour le test de comparaison est définie par :

$$F = \frac{n_1 + n_2 - (p+1)}{(n_1 + n_2 - 2)p} T^2$$

Cette statistique suit une loi de Fisher à p et $n1 + n2 - p - 1$ degrés de liberté si les échantillons suivent une loi normale pour les différentes variables.

Remarques :

- Ce test ne peut être utilisé que si l'on fait l'hypothèse que les variables sont distribuées suivant une loi normale. Il conviendra de valider cette hypothèse avec des tests de normalité.

- L'utilisation de ce test suppose que les matrices de covariance des deux groupes sont identiques, hypothèse qu'il conviendra de valider avec les tests de Box.

- Si l'on souhaite comparer plus de deux échantillons, le test basé sur la distance de Mahalanobis peut être utilisé pour identifier les sources possibles de la différence observée au niveau global. Il est alors recommandé de faire la correction de Bonferroni pour le seuil alpha. Pour k échantillons à comparer deux à deux, on utilise alors le niveau alpha suivant :

$$\alpha^* = \frac{2\alpha}{k(k-1)}$$

Lambda de Wilks

Le lambda de Wilks est une statistique dont la particularité est de suivre la distribution de Wilks à trois paramètres, définie par le rapport suivant :

$$\Lambda(p, m, n) = \frac{|A|}{|A + B|}$$

où A et B sont deux matrices semi-définies positives qui elles-mêmes suivent respectivement une loi de Wishart $W_p(I, m)$ et $W_p(I, n)$, où I est la matrice identité.

Lorsque l'on veut comparer les moyennes de p variables pour k groupes (ou échantillons ou classes) indépendants, en posant pour hypothèse nulle H0 que les p moyennes sont égales, si on suppose que les matrices de covariance sont identiques pour les k groupes, alors on montre que le test de l'hypothèse H0 revient à calculer la statistique

$$\Lambda(p, n - k, k - 1) = \frac{|W|}{|W + B|}$$

où

- W est la matrice de covariance intra-classe combinée,

- B est la matrice de covariance inter-classes,

- n est le nombre total d'observations.

La loi du lambda de Wilks étant complexe, on utilise la statistique F de Rao donnée par :

$$F = \frac{(1-\Lambda^{1/s})}{\Lambda^{1/s}} \frac{m_2}{m_1}$$

avec

$$\begin{aligned} s &= \sqrt{\frac{p^2(k-1)^2-4}{p^2+(k-1)^2-5}} \\ m_1 &= p(k-1) \\ m_1 &= s[n-(p+k+2)/2] - p(k-1)/2 + 1 \end{aligned}$$

On montre que si la taille de l'échantillon est grande, F suit une loi de Fisher à $m1$ et $m2$ degrés de liberté. Lorsque $p \leq 2$ ou $k = 2$, la statistique F suit exactement une loi de Fisher.

Remarques :

- Ce test ne peut être utilisé que si l'on fait l'hypothèse que les p variables sont distribuées suivant une loi normale. Il conviendra de valider cette hypothèse avec des tests de normalité.

- L'utilisation de ce test suppose que les matrices de covariance des k groupes sont identiques, hypothèse qu'il conviendra de valider avec les tests de Box.

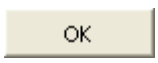
Test d'égalité des matrices de covariance intra-classe

Test de Box : le test de Box permet de tester l'hypothèse d'égalité des matrices de covariance intra-classe. Deux approximations ont été proposées, l'une basée sur la distribution du χ^2, l'autre sur la distribution de Fisher.

Test de Kullback : le test de Kullback permet de tester l'hypothèse d'égalité des matrices de covariance intra-classe. La statistique calculée est approximativement distribuée suivant une loi du χ^2.

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

Annuler : cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

Aide : cliquez sur ce bouton pour afficher l'aide.

 : cliquez sur ce bouton pour rétablir les options par défaut.

 : cliquez sur ce bouton pour effacer les sélections de données.

 : cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.

Onglet **Général**:

Tableau observations/variables : sélectionnez un tableau comprenant N objets décrits par P descripteurs. Si des en-têtes de colonnes ont été sélectionnés, veuillez vérifier que l'option « Libellés des variables » est activée. Si plusieurs variables sont sélectionnées, elles seront chacune à leur tour discrétisées.

Groupes : activez cette option pour sélectionner les données qui correspondent à l'identifiant du groupe auquel appartient chaque observation.

Poids : activez cette option si vous voulez pondérer les observations. Si vous n'activez pas cette option, les poids seront tous considérés comme valant 1. Les poids doivent être impérativement supérieurs ou égaux à 0. Si un en- tête de colonne a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des variables : activez cette option si la première ligne des données sélectionnées contient un libellé.

Onglet **Options**:

Test du lambda de Wilks : activez cette option pour que XLSTAT calcule la statistique Lambda et la p-value associée.

Test de Mahalanobis : activez cette option pour que XLSTAT calcule les distances de Mahalanobis, ainsi que les statistiques F et les p-value associées.

- **Correction de Bonferroni** : activez cette option si vous souhaitez utiliser une correction de Bonferroni lors du calcul des p-values associées aux distances de Mahalanobis.

Test de Box : activez cette option pour que XLSTAT effectue le test de Box et les p-value découlant des deux approximations possibles.

Test de Kullback : activez cette option pour que XLSTAT effectue le test de Kullback.

Niveau de signification (%) : entrez le niveau de signification à utiliser pour les différents tests (valeur par défaut : 5%).

Onglet **Données manquantes** :

Ne pas accepter les valeurs manquantes : activez cette option pour que XLSTAT empêche la poursuite des calculs si des valeurs manquantes sont détectées.

Supprimer les observations : activez cette option pour supprimer les observations comportant des données manquantes.

Onglet **Sorties**:

Statistiques descriptives : activez cette option pour afficher les statistiques descriptives pour les variables sélectionnées.

Corrélations : activez cette option pour afficher la matrice de corrélation.

Matrices de covariance : activez cette option pour afficher les matrices de covariance inter-classes, intra-classe, intra-classe totale, et totale.

Résultats

Les résultats affichés par XLSTAT correspondent aux différents tests sélectionnés.

Exemple

Un exemple de test multidimensionnel est disponible sur le Centre d'aide XLSTAT :

http://www.xlstat.com/demo-mahaf.htm

Bibliographie

Jobson J.D. (1992). Applied Multivariate Data Analysis. Volume II: Categorical and Multivariate Methods. Springer-Verlag, New York.

Legendre P. and Legendre L. (1998). Numerical Ecology. Second English Edition. Elsevier, Amsterdam.

Test z pour une proportion

Utilisez cet outil pour comparer une proportion calculée à partir d'un échantillon à une proportion donnée.

Dans cette section :

Description

Soit n le nombre d'observations vérifiant une certaine propriété parmi un échantillon de taille N . On définit par $p = n/N$, la proportion de l'échantillon vérifiant la propriété. Soit p_0 une proportion connue à laquelle on veut comparer p. Soit D la différence (exacte, minimale ou maximale) supposée entre les deux proportions. Classiquement, D est fixée à 0.

Le test bilatéral correspond au test de la différence entre $p - p_0$ et D , et les hypothèses nulle (H0) et alternative (Ha) sont les suivantes :

- H0: $p - p_0 = D$
- Ha: $p - p_0 \neq D$

Dans le cas unilatéral, il faut distinguer le test unilatéral à gauche (ou inférieur) et le test unilatéral à droite (ou supérieur). Dans le test unilatéral à gauche, les hypothèses sont les suivantes :

- H0: $p - p_0 = D$
- Ha: $p - p_0 < D$

Pour le test unilatéral à droite, les hypothèses sont les suivantes :

- H0: $p - p_0 = D$
- Ha: $p - p_0 > D$

Ce test a été développé en considérant que :

- les observations sont mutuellement indépendantes,
- la probabilité p de posséder la propriété est la même pour toutes les observations,
- l'effectif N est assez grand, et p n'est pas trop proche de 0 ou de 1.

Remarque : une règle simple pour déterminer si N est assez grand, consiste à vérifier que :

$$\begin{cases} 0 < p - 2\sqrt{p(1-p)/N} \\ p + 2\sqrt{p(1-p)/N} < 1 \end{cases}$$

Statistique z

On trouve plusieurs façons de calculer la statistique z dans la littérature statistique. La plus commune est :

$$z = \frac{p - p_0 - D}{\sigma}$$

L'approximation pour les échantillons où N est grand consiste à estimer la variance de par :

$$\hat{\sigma^2}(z) = \sqrt{\frac{p(1-p)}{N}}$$

Toutefois, si l'on pense que la que la proportion p_0 est un meilleur estimateur, on peut utiliser :

$$\hat{\sigma^2}(\pi) = \sqrt{\frac{p_0(1-p_0)}{N}}$$

Cette version ne doit toute fois pas être utilisée dans le cas où $D = 0$.

La statistique z statistic est distributée aymptotiquement suivant une loi normale. Plus N est grand, meilleure est l'approximation. La p-valeur est calculée suivant l'approximation normale.

Intervalles de confiance

Il existe plusieurs méthodes pour calculer les intervalles de confiance sur une proportion. XLSTAT propose le choix entre quatre versions différentes: Wald, Wilson score, Clopper-Pearson, Agresti Coull.

Boîte de dialogue

OK : cliquez sur ce bouton pour lancer les calculs.

Annuler : cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

Aide : cliquez sur ce bouton pour afficher l'aide.

 : cliquez sur ce bouton pour rétablir les options par défaut.

 : cliquez sur ce bouton pour effacer les sélections de données.

Onglet **Général**:

Effectif / Proportion : entrez la valeur de l'effectif n pour lequel la propriété est observée (voir la section description), ou la proportion correspondante (voir « format de données », ci-dessous).

Taille d'échantillon : entrez le nombre d'observations de l'échantillon.

Proportion test : entrez la valeur de la proportion test à laquelle la proportion observée doit être comparée.

Format des données : choisissez ici si vous préférez entrer la valeur de l'**effectif** pour lequel la propriété est observée, ou la **proportion** observée.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Onglet **Options**:

Hypothèse alternative : choisissez l'hypothèse alternative à utiliser pour le test (voir la section description).

Différence supposée : entrez la valeur de la différence supposée entre les proportions.

Niveau de signification (%) : entrez la valeur du niveau de signification pour le test (valeur par défaut : 5%).

Variance : choisissez la méthode de calcul de la variance (utilisé uniquement pour le calcul de l'intervalle de confiance de Wald).

Intervalle de confiance : choisissez la méthode de calcul de l'intervalle de confiance autour de la proportion.

Résultats

Les résultats affichés par XLSTAT correspondent aux différentes statistiques du test z (différence observée, z observé, z critique, p-value, alpha), et à l'interprétation qui en découle.

Exemple

Un exemple de test de comparaison de proportions est disponible sur le Centre d'aide XLSTAT :

http://www.xlstat.com/demo-propf.htm

Bibliographie

Agresti A., and Coull B.A. (1998). Approximate is better than "exact" for interval estimation of binomial proportions. *The American Statistician*, **52**, 119-126.

Clopper C.J. and Pearson E.S. (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26, 404-413.

Fleiss J.L. (1981). Statistical Methods for Rates and Proportions. John Wiley & Sons, New York.

Sincich T. (1996). Business Statistics by Example, 5th Edition. Prentice- Hall, Upper Saddle River.

Sokal R.R. & Rohlf F.J. (1995). Biometry. The Principles and Practice of Statistics in Biological Research. Third Edition. Freeman, New York.

Wilson, E.B. (1927). Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, **22**, 209-212.

Wald, A., & Wolfowitz, J. (1939). Confidence limits for continuous distribution functions. The Annals of Mathematical Statistics, 10, 105-118.

Test z pour deux proportions

Utilisez cet outil pour comparer deux proportions calculées pour deux échantillons.

Dans cette section :

Description

Soit n_1 le nombre d'observations vérifiant une certaine propriété pour un échantillon E_1 de taille N_1, et n_2 le nombre d'observations vérifiant la même propriété pour un échantillon E_2 de taille N_2. On définit par $p_1 = n_1/N_1$, la proportion de l'échantillon E1 vérifiant la propriété, et par $p_2 = n_2/N_2$ la proportion pour E2. Soit D la différence (exacte, minimale ou maximale) supposée entre les deux proportions. Classiquement, D est fixée à 0.

Le test bilatéral correspond au test de la différence entre $p_1 - p_2$ et D, et les hypothèses nulle (H0) et alternative (Ha) sont les suivantes :

- H0: $p_1 - p_2 = D$
- Ha: $p_1 - p_2 \neq D$

Dans le cas unilatéral, il faut distinguer le test unilatéral à gauche (ou inférieur) et le test unilatéral à droite (ou supérieur). Dans le test unilatéral à gauche, les hypothèses sont les suivantes :

- H0: $p_1 - p_2 = D$
- Ha: $p_1 - p_2 < D$

Pour le test unilatéral à droite, les hypothèses sont les suivantes :

- H0: $p_1 - p_2 = D$
- Ha: $p_1 - p_2 > D$

Ce test a été développé en considérant que :

- les observations sont mutuellement indépendantes,
- la probabilité p_1 de posséder la propriété est la même pour toutes les observations de l'échantillon E_1,
- la probabilité p_2 de posséder la propriété est la même pour toutes les observations de l'échantillon E_2,
- les effectifs N_1 et N_2 sont assez grands, et p_1 et p_2 ne sont pas trop proches de 0 ou de 1.

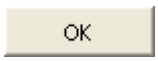
Remarque : une règle simple pour déterminer si N_1 et N_2 sont assez grands, consiste à vérifier que :

$$\begin{cases} 0 < p_1 - 2\sqrt{p_1(1-p_1)/N_1} \\ p_1 + 2\sqrt{p_1(1-p_1)/N_1} < 1 \end{cases}$$

et

$$\begin{cases} 0 < p_2 - 2\sqrt{p_2(1-p_2)/N_2} \\ p_2 + 2\sqrt{p_2(1-p_2)/N_2} < 1 \end{cases}$$

Boîte de dialogue



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.

Onglet **Général**:

Effectif 1 / Proportion 1 : entrez la valeur de l'effectif *n1* pour lequel la propriété est observée (voir la section description), ou la proportion correspondante (voir « format de données », ci-dessous).

Taille d'échantillon 1 : entrez le nombre d'observations de l'échantillon 1.

Effectif 2 / Proportion 2 : entrez la valeur de l'effectif *n2* pour lequel la propriété est observée (voir la section description), ou la proportion correspondante (voir « format de données », ci-dessous).

Taille d'échantillon 2 : entrez le nombre d'observations de l'échantillon 2.

Format des données : choisissez ici si vous préférez entrer la valeur des **effectifs** pour lesquels la propriété est observée, ou les **proportions** observées.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Test z : activez cette option pour utiliser le test z.

Méthode Monte Carlo : activez cette option pour utiliser la méthode par simulations et entrez alors le nombre de simulations à effectuer.

Onglet **Options**:

Hypothèse alternative : choisissez l'hypothèse alternative à utiliser pour le test (voir la section description).

Différence supposée : entrez la valeur de la différence supposée entre les proportions.

Niveau de signification (%) : entrez la valeur du niveau de signification pour le test (valeur par défaut : 5%).

Variance : choisissez la méthode de calcul de la variance de la différence entre les proportions.

- **p1q1/n1+p2q2/n2** : activez cette option pour calculer la variance selon cette formule.
- **pq(1/n1+1/n2)** : activez cette option pour calculer la variance selon cette formule.

Résultats

Les résultats affichés par XLSTAT correspondent aux différentes statistiques du test z (différence observée, z observé, z critique, p-value, alpha), et à l'interprétation qui en découle.

Exemple

Un exemple de test de comparaison de proportions est disponible sur le Centre d'aide XLSTAT :

http://www.xlstat.com/demo-prop2f.htm

Bibliographie

Fleiss J.L. (1981). Statistical Methods for Rates and Proportions. John Wiley and Sons, New York.

Sincich T. (1996). Business Statistics by Example, 5th Edition. Prentice- Hall, Upper Saddle River.

Comparaison de k proportions

Utilisez cet outil pour comparer k proportions et pour déterminer si elles peuvent être considérées comme égales, ou si au moins 2 proportions parmi les k sont significativement différentes.

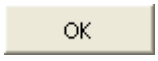
Dans cette section :

Description

XLSTAT propose trois approches différentes pour déterminer si k proportions peuvent être considérées comme étant toutes égales (hypothèse nulle H0) ou si au moins deux proportions sont différentes (hypothèse alternative Ha) :

- **Test du Khi² :** Ce test est identique à celui utilisé pour les tableaux de contingence ;
- **Méthode Monte-Carlo :** La méthode Monte Carlo permet de calculer une distribution de la distance du Khi² sur la base de simulations ayant pour contrainte de respecter les effectifs totaux pour les k groupes. On obtient donc une distribution empirique donnant une valeur critique plus fiable (à condition que le nombre de simulations soit important) que celle donnée par la distribution théorique du Khi² qui correspond au cas asymptotique.
- **Procédure de Marascuilo :** Il est conseillé de n'utiliser la procédure de Marascuilo que si le test du Khi² ou si le test équivalent faisant intervenir des simulations de Monte Carlo ont rejeté H0. La procédure de Marascuilo consiste à effectuer des tests de comparaison deux à deux pour tous les couples de proportions, ce qui permet d'identifier quelles sont les proportions responsables de l'éventuel rejet de H0.

Boîte de dialogue



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.

 : cliquez sur ce bouton pour rétablir les options par défaut.

 : cliquez sur ce bouton pour effacer les sélections de données.

Effectifs / Proportions : sélectionnez les données sur la feuille Excel.

Taille des échantillons : sélectionnez les données correspondant aux tailles des échantillons.

Libellés des échantillons : activez cette option si vous voulez utiliser des libellés d'échantillons pour l'affichage des résultats. Si l'option « Libellés des colonnes » est activée, la première cellule de la sélection doit comprendre un en-tête. Si vous n'activez pas cette option, des libellés seront automatiquement créés (Ech1, Ech 2, …).

Format des données : choisissez ici si vous préférez entrer la valeur des **effectifs** pour lesquels la propriété est observée, ou les **proportions** observées.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des colonnes : activez cette option si la première ligne des données sélectionnées (effectifs/proportions, taille des échantillons et libellés des échantillons) contient un libellé.

Test du Khi² : activez cette option pour utiliser le test du Khi².

Méthode Monte Carlo : activez cette option pour utiliser la méthode par simulations et entrez le nombre de simulations.

Procédure de Marascuilo : activez cette option pour utiliser la procédure de Marascuilo.

Niveau de signification (%) : entrez la valeur du niveau de signification pour les trois tests (valeur par défaut : 5%).

Résultats

Les résultats du test du Khi² sont affichés en premier si l'option correspondante a été activée. Pour le test du Khi² et la méthode Monte Carlo, la p-value est comparée au niveau de signification afin de valider ou non l'hypothèse nulle.

Les résultats obtenus à partir des simulations Monte Carlo seront d'autant plus proches des résultats du test du Khi² que les effectifs totaux et le nombre de simulations sont élevés. La différence se manifeste au niveau de la valeur critique et de la p-value.

La procédure de Marascuilo permet d'identifier quelles sont les proportions responsables de l'éventuel rejet de l'hypothèse nulle. Dans la colonne « Significatif » on peut identifier quelles proportions sont significativement différentes deux à deux.

Remarque : il se peut que la procédure de Marascuilo n'identifie pas de différence significative alors que le test du Khi² rejette l'hypothèse nulle. En général cela se produit lorsque deux proportions sont presque significativement différentes au niveau de la procédure de Marascuilo. Une analyse plus poussée sera alors nécessaire.

Exemple

Un exemple de test de comparaison de k proportions est disponible sur

http://www.xlstat.com/demo-kpropf.htm

Bibliographie

Agresti A. (1990). Categorical Data Analysis. John Wiley and Sons, New York.

Marascuilo L. A. and Serlin R. C. (1988). Statistical Methods for the Social and Behavioral Sciences. Freeman, New York.

Test d'ajustement multinomial

Utilisez cet outil pour vérifier si les effectifs observés pour les modalités d'une variable qualitative correspondent aux effectifs ou aux proportions attendus.

Dans cette section :

Description

Le test d'ajustement multinomial permet de vérifier si la distribution d'un échantillon correspondant à une variable qualitative (ou quantitative discrétisée) est conforme à ce que l'on attend. On parle de test d'ajustement multinomial car il est fondé sur la loi multinomiale qui est l'extension de la loi binomiale au cas il y a plus de deux modalités possibles.

Soit k le nombre de valeurs possibles (modalités) pour la variable X. On désigne par $p_1, p_2, \ldots, p_k$ les probabilités (ou densité) associées à chacune de ces valeurs.

Soit un échantillon de taille N pour lequel on obtient les effectifs suivants pour les différentes modalités : $n_1, n_2, \ldots, n_k$.

L'hypothèse nulle du test est donnée par :

- H0 : la distribution des modalités est conforme à ce que l'on attend, autrement dit la distribution de l'échantillon n'est pas différente de celle de X.

L'hypothèse alternative du test est donnée par :

- Ha : la distribution des modalités n'est pas conforme à ce que l'on attend, autrement dit la distribution de l'échantillon n'est pas identique à celle de X.

Différentes méthodes et statistiques ont été proposées pour ce test. Les possibilités offertes par XLSTAT sont :

1. Test du Khi² :

On calcule la statistique suivante :

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - Np_i)^2}{Np_i}$$

Cette statistique suit asymptotiquement une loi du Khi à k-1 degrés de liberté.

2. Test Monte Carlo :

Cette version du test permet de s'affranchir des calculs parfois lourds de la méthode exacte basée sur la loi multinomiale et d'éviter l'approximation par la loi du Khi² qui peut être de qualité moyenne sur les petits échantillons. Ce test consiste en un rééchantillonnage aléatoire de N observations dans une loi ayant les propriétés attendues. On calcule pour chaque rééchantillonnage la statistique χ^2, puis une fois ce processus terminé, on évalue combien de fois la valeur observé sur l'échantillon de départ est dépassée. On en déduit ainsi la p-value.

Boîte de dialogue

OK : cliquez sur ce bouton pour lancer les calculs.

Annuler : cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

Aide : cliquez sur ce bouton pour afficher l'aide.

 : cliquez sur ce bouton pour rétablir les options par défaut.

 : cliquez sur ce bouton pour effacer les sélections de données.

Effectifs : sélectionnez les données correspondant aux effectifs observés sur la feuille Excel.

Effectifs attendus / Proportions attendues : sélectionnez sur la feuille Excel les données correspondant aux effectifs attendus ou aux proportions attendues sur la feuille Excel.

Format des données : choisissez ici si les valeurs sélectionnées ci-dessus sont des **effectifs attendus**, ou les **proportions attendues**.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des colonnes : activez cette option si la première ligne des données sélectionnées (effectifs/proportions, taille des échantillons et libellés des échantillons) contient un libellé.

Test du Khi² : activez cette option pour utiliser le test du Khi².

Méthode Monte Carlo : activez cette option pour utiliser la méthode par simulations et entrez le nombre de simulations.

Niveau de signification (%) : entrez la valeur du niveau de signification pour les deux tests (valeur par défaut : 5%).

Résultats

Les résultats du test du Khi² sont affichés en premier si l'option correspondante a été activée. Pour le test du Khi² et la méthode Monte Carlo, la p-value est comparée au niveau de signification afin de valider ou non l'hypothèse nulle.

Les résultats obtenus à partir des simulations Monte Carlo seront d'autant plus proches des résultats du test du Khi² que les effectifs totaux et le nombre de simulations sont élevés. La différence se manifeste au niveau de la valeur critique et de la p-value.

Pour les simulations Monte Carlo un intervalle de confiance autour de la p-value est fourni.

Exemple

Un exemple de test d'ajustement multinomial est disponible sur le Centre d'aide XLSTAT :

http://www.xlstat.com/demo-goodnessf.htm

Bibliographie

Read T.R.C. and Cressie N.A.C. (1988). Goodness-of-Fit Statistics for Discrete Multivariate Data. Springer-Verlag, New York.

Saporta G. (1990). Probabilités, Analyse des Données et Statistique. Technip, Paris.

TOST (Test d'équivalence)

Utilisez cet outil pour appliquer un test d'équivalence entre deux échantillons indépendants, distribués suivant une loi normale.

Dans cette section :

Description

Les tests d'équivalence servent à la différence des tests d'hypothèse à valider le fait qu'une différence se trouve dans un intervalle donné.

Ce type de test est surtout utilisé pour valider la bioéquivalence. Ainsi, lorsque qu'on veut montrer l'équivalence entre deux médicaments, les tests d'hypothèse classiques ne s'appliquent pas, on utilisera alors des tests d'équivalence qui permettront de valider l'équivalence entre les deux médicaments.

Dans le cadre d'un test d'hypothèse classique, on cherche à rejeter l'hypothèse nulle d'égalité. Dans le cadre d'un test d'équivalence, on cherche à valider l'équivalence entre deux échantillons. Le TOST (two one-sided test) est un test d'équivalence qui se base sur le test t classique utilisé pour tester l'hypothèse d'égalité entre deux moyennes.

On va donc avoir 2 échantillons, une différence théorique entre les moyennes ainsi qu'un intervalle dans lequel on pourra dire que les moyennes des échantillons sont équivalentes.

Le test TOST est dit paramétrique car il suppose que les échantillons sont distribués suivant des lois normales. Cette hypothèse pourra être testée à l'aide des tests de normalité.

Le test TOST utilise des tests de Student afin de vérifier l'équivalence entre les moyennes de deux échantillons. Une description détaillée de ces tests peut être trouvée dans le chapitre qui leur est consacré.

XLSTAT propose deux méthodes alternatives afin de tester l'équivalence à l'aide du test TOST.

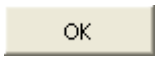
- Utilisation de l'intervalle confiance à 100*(1-2*alpha)% autour de la moyenne. En comparant cet intervalle à l'intervalle d'équivalence défini par l'utilisateur, on peut conclure à l'équivalence ou à la non-équivalence. Ainsi, si l'intervalle de confiance est compris dans l'intervalle défini par l'utilisateur, on conclut à l'équivalence entre les deux échantillons. Si l'une des bornes de l'intervalle de confiance se trouve à l'extérieur de l'intervalle défini par l'utilisateur, alors les deux échantillons ne sont pas équivalents.

- Utilisation de deux tests unilatéraux à droite et à gauche. Ainsi on applique un test t unilatéral à droite sur la borne inférieure de l'intervalle défini par l'utilisateur et un test unilatéral à gauche sur la borne supérieure de l'intervalle défini par l'utilisateur. On obtient ainsi des p-valeurs pour les deux tests. On prendra la plus grande de ces p-valeurs comme p-valeur du test d'équivalence.

Ces deux tests sont similaires et doivent donner des résultats concordants. Ils ont été introduits par Schuirman's (1987).

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.

Onglet **Général**:

Echantillon 1 : sélectionnez une colonne de données correspondant au premier échantillon.

Echantillon 2 : sélectionnez une colonne de données correspondant au second échantillon.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des colonnes/lignes : activez cette option si la première ligne (mode colonnes) ou colonne (mode lignes) des données sélectionnées contient des libellés.

Onglet **Options**:

Différence supposée (D) : entrez la valeur de la différence supposée entre les moyennes des échantillons.

Borne inférieure : entrez la valeur de la borne inférieure pour l'intervalle associé à la différence entre les moyennes des échantillons.

Borne supérieure : entrez la valeur de la borne supérieure pour l'intervalle associé à la différence entre les moyennes des échantillons.

Niveau de signification (%) : entrez la valeur du niveau de signification pour les tests (valeur par défaut : 5%).

Poids : activez cette option si vous voulez pondérer les observations. Cette option n'est visible que si vous avez choisi le format de données « une colonne par variable » ou un test apparié. Si vous n'activez pas cette option, les poids seront tous considérés comme valant 1. XLSTAT prend en compte ces poids pour les calculs des degrés de libertés. Les poids doivent être impérativement supérieurs ou égaux à 0. Si un en-tête de colonne a été sélectionné, veuillez vérifier que l'option « Libellés des colonnes/lignes » est activée.

Variances des échantillons pour le test t :

- **Supposer l'égalité** : activez cette option pour considérer que la variance des échantillons est égale.
- **Cochran-Cox** : activez cette option pour calculer la p-value en utilisant la méthode de Cochran et Cox dans le cas où les variances ne sont pas supposées égales.
- **Utiliser un test F** : activez cette option pour utiliser le test F de Fisher afin de déterminer si les variances des deux échantillons peuvent être considérées comme étant égales ou non.

Onglet **Données manquantes** :

Ne pas accepter les valeurs manquantes : activez cette option pour que XLSTAT empêche la poursuite des calculs si des valeurs manquantes sont détectées.

Supprimer les observations : activez cette option pour supprimer les observations comportant des données manquantes.

Onglet **Sorties**:

Statistiques descriptives : activez cette option pour afficher les statistiques descriptives des échantillons.

Résultats

Les résultats affichés par XLSTAT correspondent aux différentes statistiques des tests sélectionnés, et à l'interprétation qui en découle.

Le premier tableau rassemble les statistiques descriptives associées aux deux échantillons.

Le tableau de résultats suivant permet de valider l'hypothèse d'équivalence, pour que deux moyennes soient équivalentes, il faut que l'intervalle de confiance autour de la différence avec un niveau de confiance de 1-2*alpha soit compris dans l'intervalle défini par l'utilisateur dans la boîte de dialogue. Si tel est le cas, l'interprétation conduit à l'équivalence entre les moyennes. Il faut regarder si les 4 valeurs de ce tableau sont bien ordonnées de manière croissante. La dernière ligne donne une interprétation (équivalence ou non-équivalence).

Le tableau suivant permet de visualiser les deux tests unilatéraux à gauche et à droite en se basant sur les bornes définies par l'utilisateur. Il s'agit d'une autre interprétation du test d'équivalence. La p-valeur pour le test d'équivalence est la plus grande p-valeur unilatérale obtenue.

Exemple

Un exemple de test d'équivalence (TOST) pour deux échantillons est disponible sur

http://www.xlstat.com/demo-tostf.htm

Bibliographie

Satterthwaite F.W. (1946). An approximate distribution of estimates of variance components. *Biometrics Bulletin*, **2**, 110 -114.

Schuirmann, D.J. (1987). A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics*, **15**, 657–680.

Sokal R.R. and Rohlf F.J. (1995). Biometry. The Principles and Practice of Statistics in Biological Research. Third Edition. Freeman, New York.

Tomassone R., Dervin C. and Masson J.P. (1993). Biométrie. Modélisation de Phénomènes Biologiques. Masson, Paris.

Tests non paramétriques

Comparaison de deux distributions (Kolmogorov-Smirnov)

Utilisez cet outil pour comparer les distributions de deux échantillons et pour déterminer si elles peuvent être considérées comme identiques.

Dans cette section :

Description

Le test de Kolmogorov-Smirnov permet de comparer deux distributions. Ce test est utilisé pour les tests d'ajustement d'une distribution pour comparer une distribution empirique déterminée à partir d'un échantillon à une distribution connue. Il peut aussi être utilisé pour comparer deux distributions empiriques.

Remarque : ce test permet de tester l'identité des distributions, à la fois quant à leur forme et à leur position.

Soit un échantillon E_1, comprenant n_1 observations, et F_1 la fonction de répartition empirique correspondante. Soit un second échantillon E_2, comprenant n_2 observations, et F_2 la fonction de répartition empirique correspondante.

L'hypothèse nulle du test de Kolmogorov-Smirnov est définie par :

H0: $F_1(x) = F_2(x)$

La statistique de Kolmogorov est définie par :

$$D_1 = \sup_x |F_1(x) - F_2(x)|$$

D_1 est la différence absolue maximale entre les deux distributions empiriques. Sa valeur est donc comprise entre 0 (cas d'une identité parfaite des distributions) et 1 (cas d'une séparation parfaite des distributions). L'hypothèse alternative associée à cette statistique est :

Ha: $F_1(x) \neq F_2(x)$

Les statistiques de Smirnov sont définies par :

$$D_2 = \sup_x \{F_1(x) - F_2(x)\}$$

$$D_3 = \sup_x \{F_2(x) - F_1(x)\}$$

L'hypothèse alternative associée à D_2 est :

Ha: $F_1(x) < F_2(x)$

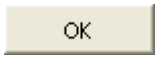
L'hypothèse alternative associée à D_3 est :

Ha: $F_1(x) > F_2(x)$

Nikoforov (1994) a proposé une méthode de test exact pour le test de Kolmogorov-Smirnov sur deux échantillons. Cette méthode est utilisée par XLSTAT pour les trois hypothèses alternatives. XLSTAT permet aussi d'introduire la différence D supposée entre les distributions. Cette valeur doit être comprise entre 0 et 1.

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.

Onglet **Général**:

Données / Echantillon 1 : si le format de données sélectionné est « une colonne par variable », sélectionnez les données correspondant aux différents échantillons sur la feuille Excel. Si le format de données sélectionné est « une colonne par échantillon », sélectionnez une colonne de données correspondant au premier échantillon.

Identifiant d'échantillon / Echantillon 2 : si le format de données sélectionné est « une colonne par variable», sélectionnez les données identifiant les deux échantillons auxquels les données sélectionnées correspondent. Si le format de données sélectionné est « une colonne par échantillon » sélectionnez une colonne de données correspondant au second échantillon.

Format des données : choisissez le format des données.

- **Une colonne/ligne par échantillon** : activez cette option pour sélectionner une colonne (ou ligne en mode lignes) par échantillon.
- **Une colonne/ligne par variable** : activez cette option pour que XLSTAT fasse autant de tests qu'il y a de colonnes/lignes, sachant que chaque colonne/ligne doit contenir le même nombre de lignes/colonnes, et qu'un identifiant d'échantillon permettant d'affecter chaque observation à un échantillon doit par ailleurs être sélectionné.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des colonnes/lignes : activez cette option si la première ligne (mode colonnes) ou colonne (mode lignes) des données sélectionnées contient des libellés.

Test de Kolmogorov-Smirnov : activez cette option pour utiliser le test de Kolmogorov-Smirnov (voir description).

Onglet **Options**:

Hypothèse alternative : choisissez l'hypothèse alternative à utiliser pour le test (voir description).

Différence supposé e (D) : entrez la valeur de la différence maximale supposée entre les fonctions de répartition empiriques des échantillons. La différence doit être comprise entre 0 et 1.

Niveau de signification (%) : entrez la valeur du niveau de signification pour les tests (valeur par défaut : 5%).

Onglet **Données manquantes** :

Ne pas accepter les valeurs manquantes : activez cette option pour que XLSTAT empêche la poursuite des calculs si des valeurs manquantes sont détectées.

Supprimer les observations : activez cette option pour supprimer les observations comportant des données manquantes.

Onglet **Sorties**:

Statistiques descriptives : activez cette option pour afficher les statistiques descriptives des échantillons.

Onglet **Graphiques** :

Diagramme de dominance : activez cette option pour afficher un diagramme de dominance afin de comparer visuellement les échantillons.

Histogrammes cumulés : activez cette option pour afficher un graphique permettant de visualiser les fonctions de répartition empiriques des échantillons.

Résultats

Les résultats affichés par XLSTAT correspondent aux différentes statistiques des tests sélectionnés, et à l'interprétation qui en découle.

Exemple

Bibliographie

Abramowitz M. and Stegun I.A. (1972). Handbook of Mathematical Functions. Dover Publications, New York.

Durbin J. (1973). Distribution Theory for Tests Based on the Sample Distribution Function. SIAM, Philadelphia.

Kolmogorov A. (1941). Confidence limits for an unknown distribution function. *Ann. Math. Stat.* **12**, 461–463.

Nikiforov A.M. (1994). Algorithm AS 288: Exact two-sample Smirnov test for arbitrary distributions. *Applied .s tatistics*, 43(1), 265-270.

Smirnov N. V. (1939). On the estimation of the discrepancy between empirical curves of distribution for two independent samples. *Bulletin Moscow University*, **2**, 3-14.

Tests des médianes (test de Mood)

Utilisez cet outil pour tester si k échantillons indépendants ont la même médiane.

Dans cette section :

Description

Le test de Mood (ou le test de la médiane) a été proposé en 1950 pour déterminer si k échantillons ($k \geq 2$) ont la même médiane. Il s'agit d'un test non paramétrique (ne faisant donc pas d'hypothèse sur la distribution des mesures faites) et peut être vu comme un cas particulier du test du Khi² de Pearson.

Test de Mood :

On désigne par Mi la médiane de l'échantillon i, l'hypothèse nulle H0 et alternative Ha du test de Mood sont les suivantes :

- H0 : M1 = M2 = … = Mk
- Ha : Il existe au moins un couple (i,j) tel que Mi $\neq$ Mj

La statistique U du test de Mood est obtenue à partir d'un tableau de contingence (2 x k) suivant :

Echantillons	1	2	…	k	Total
$>$**Médiane**	O_{11}	O_{12}	…	O_{1k}	a
$\leq$**Médiane**	O_{21}	O_{22}	…	O_{2k}	b
Total	n_1	n_2	…	n_k	N

Si XLSTAT détecte un trop grand nombre d'égalités avec la médiane, les observations égales à la médiane seront automatiquement comptabilisées dans le groupe des observations qui lui sont strictement supérieures. Ceci afin de rendre la statistique ci-dessous calculable.

La statistique s'écrit

$$U = \frac{N^2}{ab} \sum_{i=1}^{k} \frac{(O_{1i} - \frac{n_i a}{N})}{n_i}$$

où N, n, a, b, et O sont définis dans le tableau de contingence.

Cette statistique a la propriété d'être asymptotiquement distribuée suivant une loi du Khi² à k-1 degrés de liberté. Yates (1934) a proposé une correction de continuité lorsque le nombre d'échantillon k est 2. On note U_Y cette statistique.

$$U_Y = \frac{N(|O_{11}O_{22} - O_{12}O_{21}| - N/2)^2}{n_1 n_2 ab}$$

Calcul de la p-value

Pour le calcul de la p-value associée à la statistique U, XLSTAT propose trois alternatives :

- Méthode asymptotique : la p-value est obtenue grâce à une approximation de la loi de U par un Khi² à (k-1) degré de liberté. Cette approximation est d'autant plus fiable que le nombre d'observation est important.
- Méthode Monte Carlo : ce calcul est basé sur un rééchantillonnage aléatoire. L'utilisateur doit choisir le nombre de simulations (ou rééchantillonnages) à réaliser. Un intervalle de confiance autour de la p-value obtenue est fourni. Cet intervalle sera bien entendu d'autant plus resserré que le nombre de simulations est important.
- Méthode exacte : lorsque le nombre d'échantillons k est 2, la probabilité d'obtenir une certaine configuration dans le tableau de contingence s'obtient directement à partir de la loi hypergéométrique. Ce calcul est d'autant plus intéressant lorsque le nombre d'observations est petit, c'est-à-dire quand l'approximation par un Khi² n'est pas satisfaisante.

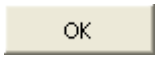
Afin d'éviter un blocage d'Excel dans le cas de la méthode Monte Carlo, XLSTAT donne la possibilité à l'utilisateur de fixer le temps maximum, exprimé en secondes, qu'il souhaite consacrer à la recherche de la p-value.

Comparaisons multiples par paires

Si la p-value est telle que l'on doit rejeter l'hypothèse H0, alors au moins une variable a une médiane différente des autres. Afin d'identifier quel(s) variable(s) est/sont responsable(s) du rejet de H0, il est possible d'utiliser une procédure de comparaisons multiples.

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas (mode colonnes), XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite (mode lignes), XLSTAT considère que les variables sont en lignes et les observations en colonnes.

Onglet **Général**:

Tableau Individus/Variables : sélectionnez un tableau dont les lignes correspondent aux individus et les colonnes aux variables (ou vice versa si le mode lignes est activé). Si les libellés des variables ont été sélectionnés, veillez à ce que l'option « libellés des colonnes » soit activée.

Format des données : choisissez le format des données.

- Une colonne/ligne par échantillon : activez cette option pour sélectionner une colonne (ou ligne en mode lignes) par échantillon.
- Une colonne/ligne par variable : activez cette option pour que XLSTAT fasse autant de tests qu'il y a de colonnes/lignes, sachant que chaque colonne/ligne doit contenir le même nombre de lignes/colonnes, et qu'un identifiant d'échantillon permettant d'affecter chaque observation à un échantillon doit par ailleurs être sélectionné.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des colonnes : activez cette option si des en-têtes de colonne ont été sélectionnés (ou des en-têtes de ligne en mode lignes).

Comparaisons multiples par paires : activez cette option pour calculer les tests de comparaisons multiples par paires.

Onglet **Options**:

Niveau de signification (%) : entrez le niveau de signification à utiliser pour les différents tests (valeur par défaut : 5%).

Pour le calcul de la p-value, vous pouvez choisir pour entre la méthode asymptotique, la méthode exactes ou la méthode de Monte Carlo (voir la section description). Dans le cas de la méthode Monte Carlo, vous pouvez préciser le nombre de rééchantillonnages, et quel temps maximum vous souhaitez consacrer à l'estimation de la p-value.

Dans certain cas, au lieu de calculer la statistique classique U du test de Mood, il est possible de prendre en compte la correction de continuité de Yates (voir la section description).

Onglet **Sorties**:

Statistiques descriptives : activez cette option pour calculer et afficher les statistiques descriptives pour les différentes variables.

Résultats

Statistiques descriptives : dans ce tableau sont affichées les statistiques descriptives correspondant aux différentes variables.

Test de Mood : Les résultats qui correspondent au test de Mood sont ensuite affichés. Une interprétation du test est fournie, suivie des comparaisons multiples afin d'identifier les variables responsables du rejet de l'hypothèse nulle si elle a été rejetée.

Exemple

Un exemple de test de Mood est disponible sur le Centre d'aide XLSTAT :

http://www.xlstat.com/demo-moodf.htm

Bibliographie

Conover W.J. (1999). Practical Nonparametric Statistics, 3rd edition, *Wiley*.

Yates F. (1934). Contingency table involving small numbers and the χ^2 test. *Journal of the Royal Statistical Society*, **1(2)**, 217-235

Test des rangs signés de Wilcoxon pour un échantillon

Utilisez cet outil pour tester l'hypothèse que le paramètre de position d'un échantillon est égal à une valeur donnée.

Dans cette section :

Description

Ce test est la version non paramétrique du test de Student pour un échantillon. Il est basé sur les rangs et, pour cette raison, le paramètre de position n'est pas ici la moyenne, mais la médiane. Vous devez utiliser ce test dès lors que vous avez des doutes quant à la normalité de votre échantillon, hypothèse nécessaire pour utiliser le test de Student :

Pour le test bilatéral, les hypothèses nulle H_0 et alternative H_a sont les suivantes :

- H_0 : $Médiane = m$
- H_a : $Médiane \neq m$

Pour le test unilatéral à gauche, les hypothèses sont les suivantes :

- H_0 : $Médiane = m$
- H_a : $Médiane < m$

Pour le test unilatéral à droite, les hypothèses sont les suivantes :

- H_0 : $Médiane = m$
- H_a : $Médiane > m$

Test des rangs signés de Wilcoxon

Wilcoxon a proposé un test qui prend en compte l'importance des différences entre de paires. Ce test test est appelé test des rangs signés de Wilcoxon, car les données sont transformées en rangs et le signe des différences est pris en compte.

Pour chaque observation de l'échantillon $(X_1, X_2, \ldots, X_n)$, on calcule la différence entre la valeur observée et la médiane m saisie. Ensuite les différences sont triées en ordre croissant et on calcule leur rang, puis on signe les rangs en fonction du signe de la différence. Soient $S_1, S_2, \ldots, S_p$ les rangs signés positifs.

La statistique utilisée pour tester si la médiane de l'échantillon est égale à m est :

$$V_s = \sum_{i=1}^{p} S_i$$

L'espérance et la variance de Vs sont respectivement :

$$E(V_s) = \frac{n(n+1)}{4}$$

et

$$V(V_s) = \frac{n(n+1)(2n+1)}{24}$$

S'il y a des ex aequo parmi les différences ou des différences nulles, on a :

$$E(V_s) = \frac{n(n+1) - d_0(d_0+1)}{4}$$

$$V(V_s) = \frac{[n(n+1)(2n+1) - d_0(d_0+1)(2d_0+1)]}{24} - \frac{\sum_{i=1}^{nd}(d_i^3 - d_i)}{48}$$

où d_0 est le nombre de différences nulles, nd le nombre de différences distinctes, et d_i le nombre de valeurs correspondant à la ième différence distincte (cela revient au même que de considérer que les d_i sont les nombres d'ex aequo pour la ième valeur distincte parmi les différences).

S'il n'y a des différences nulles ou des ex aequo parmi les différences, et si $n \leq 100$, XLSTAT calcule une p-value exacte (Lehmann, 1975). Dans les autres cas, une approximation normale est utilisée :

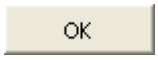
$$P(V_s \leq \nu) \approx \Phi\left(\frac{\nu - E(V_s) + c}{\sqrt{V(V_s)}}\right)$$

où f désigne la fonction de répartition de la loi normale, et c est la correction de continuité pour augmenter la qualité de l'approximation (c vaut $\frac{1}{2}$ ou $-\frac{1}{2}$ en fonction de la nature du test). Plus n est élevé, plus l'approximation est fiable.

La médiane de l'échantillon et son intervalle de confiance sont également calculés sur la base de ces calculs. Cet estimateur est fiable pour les échantillons avec une distribution symétrique. Dans le cas d'ex aequo, l'approche de Hodges et Lehmann (1963) semble toujours fiable sur la base des simulations de Monte Carlo que nous avons faites.

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.

Onglet **Général**:

Echantillons : sélectionnez les données correspondant à un ou plusieurs échantillons sur une feuille de calcul.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des colonnes/lignes : activez cette option si la première ligne (mode colonnes) ou colonne (mode lignes) des données sélectionnées contient des libellés.

Onglet **Options**:

Hypothèse alternative : choisissez l'hypothèse alternative à utiliser pour le test (voir description).

Médiane théorique (M) : entrez la valeur de la médiane supposée des échantillons.

Niveau de signification (%) : entrez la valeur du niveau de signification pour les tests (valeur par défaut : 5%).

En fonction du test utilisé, plusieurs méthodes de calcul de la p-value sont proposées. Choisissez entre la méthode **asymptotique** ou **exacte** .

Correction de continuité : activez cette option si vous souhaitez que XLSTAT utilise la correction de continuité dans le cas d'un calcul de p-value asymptotique.

Onglet **Données manquantes** :

Ne pas accepter les valeurs manquantes : activez cette option pour que XLSTAT empêche la poursuite des calculs si des valeurs manquantes sont détectées.

Supprimer les observations : activez cette option pour supprimer les observations comportant des données manquantes.

Onglet **Sorties**:

Statistiques descriptives : activez cette option pour afficher les statistiques descriptives des échantillons.

Résultats détaillés : activez cette option pour afficher les résultats détaillés des tests.

Tableau de synthèse : dans le cas où plusieurs échantillons ont été sélctionnés, activez cette option pour afficher un tableau de synthèse des p-values obtenues.

Résultats

Les résultats affichés par XLSTAT correspondent aux différentes statistiques des tests sélectionnés, et à l'interprétation qui en découle.

Exemple

Un exemple de test de Wilcoxon pour 1 échantillon est disponible sur le Centre d'aide XLSTAT

http://www.xlstat.com/demo-1-sample-wilcoxonf.htm

Bibliographie

David F. Bauer (1972). Constructing confidence sets using rank statistics. *Journal of the American Statistical Association*, **67**, 687-690.

Hollander M. and Wolfe D. A. (1999). Nonparametric Statistical Methods, Second Edition John Wiley and Sons, New York.

Hodges J. L , and Lehmann E. L. (1963). Estimation of location based on ranks. *Annals of Mathematical Statistics*, **34(2)**, 598-611.

Lehmann E.L (1975). Nonparametrics: Statistical Methods Based on Ranks. Holden-Day, San Francisco.

Wilcoxon F. (1945). Individual comparisons by ranking methods. *Biometrics*, **1**, 80-83.

Comparaison de deux échantillons (Wilcoxon, Mann-Whitney, ...)

Utilisez cet outil pour comparer deux échantillons décrits par des données quantitatives ordinales ou discrètes, qu'ils soient indépendants ou appariés.

Dans cette section :

Description

Afin de s'affranchir de l'hypothèse de normalité des échantillons nécessaire pour l'utilisation des tests paramétriques (test z, test t de Student, test F de Fisher, test de Levene, test de Bartlett), des tests non paramétriques ont été proposés.

Comme pour les tests paramétriques, on distingue le cas où les échantillons sont indépendants (par exemple, dans le cas d'une comparaison du chiffre d'affaire annuels par magasin entre deux régions pour une chaîne de supermarchés), du cas où ils sont appariés (par exemple, dans le cas d'une comparaison, à l'intérieur d'une même région, des chiffres d'affaires annuels entre deux années).

Si l'on désigne par D la différence de position supposée des échantillons (en général on teste l'égalité, et D vaut donc 0), et par $P_1 - P_2$ la différence de position des échantillons, trois types de tests sont possibles en fonction de l'hypothèse alternative choisie :

Pour le test bilatéral, les hypothèses nulle H_0 et alternative H_a sont les suivantes :

- $H_0 : P_1 - P_2 = D$
- $H_a : P_1 - P_2 \neq D$

Pour le test unilatéral à gauche, les hypothèses sont les suivantes :

- $H_0 : P_1 - P_2 = D$
- $H_a : P_1 - P_2 < D$

Pour le test unilatéral à droite, les hypothèses sont les suivantes :

- $H_0 : P_1 - P_2 = D$

- $H_a : P_1 - P_2 > D$

Comparaison de deux échantillons indépendants

Trois chercheurs, Mann, Whitney, et Wilcoxon, ont mis au point séparément un test non paramétrique très similaire qui permet de déterminer si, sur la base des rangs des échantillons, on peut considérer que les échantillons sont identiques ou non en terme de position. Ce test est souvent appelé test de **Mann-Whitney**, parfois test de Wilcoxon-Mann-Whitney, ou encore *Wilcoxon Rank-Sum test* (Lehmann, 1975).

On lit parfois que ce test permet de déterminer si les échantillons proviennent de populations ou de distributions identiques. Cela est totalement faux. Ce test permet uniquement d'étudier la position relative des échantillons. Par exemple, si on génère un échantillon de 500 observations tiré dans une loi $\mathcal{N}(0,1)$ et un échantillon de 500 observations tiré dans une loi $\mathcal{N}(0,4)$, le test de Mann-Whitney ne trouve aucune différence entre les échantillons.

Soit un échantillon E_1, comprenant n_1 observations $(x_1, x_2, \ldots, x_{n1})$ et soit E_2 un second échantillon, comprenant n_2 observations $(y_1, y_2, \ldots, y_{n1})$ et indépendant de E_1. Soit N la somme de n_1 et n_2.

Pour calculer la statistique de Wilcoxon W_s mesurant la différence de position entre le premier échantillon E_1, et l'échantillon E_2 auquel on soustrait D, on regroupe les valeurs obtenues pour les deux échantillons, puis on les ordonne. La statistique W_s est la somme des rangs de l'un des échantillons. Dans le cas de XLSTAT, la somme est calculée sur le premier échantillon.

On a alors pour l'espérance et la variance de W_s :

$$E(W_S) = \frac{1}{2}n_1(N+1) \quad \text{et} \quad V(W_S) = \frac{1}{12}n_1n_2(N+1)$$

La statistique U de Mann-Whitney est quant à elle la somme du nombre de couples (x_i, y_i) où $x_i > y_i$, parmi tous les couples possibles. On montre que :

$$E(U) = \frac{n_1n_2}{2} \text{ et} \quad V(U) = \frac{1}{12}n_1n_2(N+1)$$

On peut noter que les variances de W_s et U sont identiques. En fait, on a la relation suivante entre U et W_s :

$$W_S = U + \frac{n_1(n_1+1)}{2}$$

Les résultats proposés par XLSTAT sont ceux relatifs à la statistique U de Mann-Whitney.

Lorsqu'il y a des ex aequo parmi les valeurs des deux échantillons, le rang affecté aux valeurs ex aequo est la moyenne de leur rang avant traitement, par exemple, pour deux échantillons de taille respective, 3 et 3, si la liste des valeurs ordonnées est, {1, 1.2, 1.2, 1.4, 1.5, 1.5}, les rangs sont d'abord {1, 2, 3, 4, 5, 6} puis après prise en compte {1, 2.5, 2.5, 4, 5.5, 5.5}. Si cela ne change pas l'espérance de W_s et U, la variance est en revanche modifiée :

$$V(W_S) = V(U) = \frac{1}{12} n_1 n_2 (N+1) - \frac{n_1 n_2 \sum_{i=1}^{nd} (d_i^3 - d_i)}{12N(N-1)}$$

où nd est le nombre de valeurs distinctes, et d_i l'effectif correspondant à chacune de ces valeurs.

Pour le calcul des p-values associées à la statistique U, XLSTAT peut utiliser une méthode exacte si l'utilisateur le souhaite dans les cas suivants :

$U * n_1 * n_2 \leq 10e7$, si il n'y a pas d'ex aequo

$U * nd \leq 5000$ si il y a des ex aequo.

Les calculs peuvent être sensiblement ralentis dans le cas où il y a des ex aequo. Une approximation normale a été proposée afin de contourner ce problème. On a :

$$P(U \leq u) \approx \Phi \left(\frac{u - E(U) + c}{\sqrt{V(U)}} \right)$$

où F est la fonction de répartition de la loi normale centrée réduite, et c est une correction de continuité qui permet d'améliorer la qualité de l'approximation (c vaut $\frac{1}{2}$ ou $-\frac{1}{2}$ en fonction de la nature du test). L'approximation est d'autant plus fiable que n_1 et n_2 sont élevés.

Si l'utilisateur demande à ce qu'un test exact soit utilisé et que cela n'est pas possible en raison des contraintes énoncées ci-dessous, XLSTAT indique, dans le rapport des résultats, qu'une approximation a été utilisée.

Comparaison de deux échantillons appariés

Deux tests ont été proposés pour le cas où les échantillons sont appariés : le **test du signe** et le **test de Wilcoxon signé**.

Soit un échantillon E_1, comprenant n observations $(x_1, x_2, \ldots, x_n)$ et soit E_2 un second échantillon apparié à E_1, comprenant aussi n observations $(y_1, y_2, \ldots, y_n)$. Soit $(p_1, p_2, \ldots, p_n)$ les n paires de valeurs (x_i, y_i).

Test du signe

Soit $N+$ la statistique égale au nombre de paires telles que $y_i > x_i$, N_0 la statistique égale au nombre de paires telles que $y_i = x_i$, et $N-$ la statistique égale au nombre de paires telles que $y_i < x_i$. On montre alors que la statistique $N+$ suit une loi binomiale de paramètres $(n - N_0)$ et de probabilité $\frac{1}{2}$. L'espérance et la variance de $N+$ sont alors :

$$E(N+) = \frac{n - N_0}{2}$$

et

$$V(N+) = \frac{n - N_0}{4}$$

La p-value associée à la statistique $N+$ et au type de test choisi (bilatéral, unilatéral à droite ou unilatéral à gauche) peut donc être déterminée de manière exacte.

Remarque : ce test est appelé test du signe car il est construit à partir du signe des différences à l'intérieur des n paires. Ce test peut donc être utilisé pour comparer des évolutions évaluées sur une échelle ordinale. Par exemple, on utilisera ce test pour déterminer si l'effet d'un médicament est positif, à partir d'une enquête où le patient doit simplement déclarer s'il se sent moins bien, pas mieux, ou mieux après la prise d'un médicament.

L'inconvénient du test du signe est qu'il ne prend pas en compte l'importance de la différence entre chaque paire, information qui est pourtant souvent disponible.

Test de Wilcoxon signé

Wilcoxon a proposé un test qui permet de prendre en compte le niveau de différence à l'intérieur des paires. Ce test est appelé test de Wicoxon signé (*Wilcoxon signed rank test*), car le signe des différences intervient aussi.

Comme pour le test du signe, on calcule les différences pour l'ensemble des paires, puis on les ordonne, puis on sépare les différences positives $S_1, S_2, \ldots, S_p$ des différences négatives $R_1, R_2, \ldots, R_m$ $(p + m = n)$.

La statistique permettant de tester si les deux échantillons ont la même position ou non est définie comme la somme des S_i :

$$V_S = \sum_{i=1}^{p} S_i$$

L'espérance et la variance de V_s sont :

$$E(V_S) = \frac{n(n+1)}{4}$$

et

$$V(V_S) = \frac{n(n+1)(2n+1)}{24}$$

Dans le cas où il y aurait des ex aequo parmi les différences, ou des différences nulles pour certaines paires, on a :

$$E(V_S) = \frac{n(n+1) - d_0(d_0+1)}{4}$$

$$V(V_S) = \frac{[n(n+1)(2n+1) - d_0(d_0+1)(2d_0+1)]}{24} - \frac{\sum_{i=1}^{nd}(d_i^3 - d_i)}{48}$$

où d_0 est le nombre de différences nulles, nd le nombre de différences distinctes, et d_i l'effectif correspondant à la i-ième valeur de différence distincte (il est équivalent de considérer que les d_i est le nombre d'ex aequo pour la i-ième valeur de différence distincte).

Dans le cas où il n'y a pas de différence nulle ou d'ex aequo parmi les différences, et si $n \leq 100$, XLSTAT calcule une p-value exacte (Lehmann, 1975). Dans le cas où il y a des ex aequo, une approximation normale est utilisée. On a en effet :

$$P(V_S \leq \nu) \approx \Phi\left(\frac{\nu - E(V_S) + c}{\sqrt{V(V_S)}}\right)$$

où F est la fonction de répartition de la loi normale centrée réduite, et c est une correction de continuité qui permet d'améliorer la qualité de l'approximation (c vaut $\frac{1}{2}$ ou $-\frac{1}{2}$ en fonction de la nature du test). L'approximation est d'autant plus fiable que n est grand.

Calcul de la p-value par simulations Monte Carlo

Pour le calcul de la p-value associée à une valeur donnée de la statistique calculée, XLSTAT propose en fonction des tests trois alternatives :

- Méthode asymptotique : la p-value est obtenue grâce à une approximation asymptotique de distribution de la statistique calculée.
- Méthode exacte : le calcul de la p-value exacte repose sur la distribution réelle de la statistique calculée. Ce calcul est parfois très intensif numériquement.
- Méthode Monte Carlo : ce calcul est basé sur un rééchantillonnage aléatoire. L'utilisateur doit choisir le nombre de simulations (ou rééchantillonnages) à réaliser. Un intervalle de confiance autour de la p-value obtenue est fourni. Cet intervalle sera bien entendu d'autant plus resserré que le nombre de simulations est important. Cette approche est proposée pour le test de Mann-Whitney et pour le test de Wilcoxon signé.

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.

OK : cliquez sur ce bouton pour lancer les calculs.

Annuler : cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.

Onglet **Général**:

Données / Echantillon 1 : si le format de données sélectionné est « une colonne par variable », sélectionnez les données correspondant aux différents échantillons sur la feuille Excel. Si le format de données sélectionné est « une colonne par échantillon » ou « échantillons appariés », sélectionnez une colonne de données correspondant au premier échantillon.

Identifiant d'échantillon / Echantillon 2 : si le format de données sélectionné est « une colonne par variable», sélectionnez les données identifiant les deux échantillons auxquels les données sélectionnées correspondent. Si le format de données sélectionné est « une colonne par échantillon » ou « échantillons appariés » sélectionnez une colonne de données correspondant au second échantillon.

Format des données : choisissez le format des données.

- **Une colonne/ligne par échantillon** : activez cette option pour sélectionner une colonne (ou ligne en mode lignes) par échantillon.
- **Une colonne/ligne par variable** : activez cette option pour que XLSTAT fasse autant de tests qu'il y a de colonnes/lignes, sachant que chaque colonne/ligne doit contenir le même nombre de lignes/colonnes, et qu'un identifiant d'échantillon permettant d'affecter chaque observation à un échantillon doit par ailleurs être sélectionné.
- **Echantillons appariés** : activez cette option pour faire des tests sur échantillons appariés. Vous devez alors sélectionner une colonne (ou ligne en mode lignes) par échantillon, tout en veillant à ce que les échantillons soient de même taille.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des colonnes/lignes : activez cette option si la première ligne (mode colonnes) ou colonne (mode lignes) des données sélectionnées contient des libellés.

Test de Mann-Whitney : activez cette option pour utiliser le test de Mann-Whitney (voir description).

Test du signe : activez cette option pour utiliser le test du signe (voir description).

Test de Wilcoxon signé : activez cette option pour utiliser le test de Wilcoxon signé (voir description).

Onglet **Options**:

Hypothèse alternative : choisissez l'hypothèse alternative à utiliser pour le test (voir description).

Différence supposée (D) : entrez la valeur de la différence de position supposée entre les échantillons.

Niveau de signification (%) : entrez la valeur du niveau de signification pour les tests (valeur par défaut : 5%).

En fonction du test utilisé, plusieurs méthodes de calcul de la p-value sont proposées. Choisissez entre la méthode **asymptotique**, **exacte** ou **Monte Carlo** (voir la section description). Dans le cas des méthodes exacte et Monte Carlo, vous pouvez préciser quel temps maximum vous souhaitez consacrer à la recherche de la p-value.

Correction de continuité : activez cette option si vous souhaitez que XLSTAT utilise la correction de continuité dans le cas d'un calcul de p-value asymptotique.

Onglet **Données manquantes** :

Ne pas accepter les valeurs manquantes : activez cette option pour que XLSTAT empêche la poursuite des calculs si des valeurs manquantes sont détectées.

Supprimer les observations : activez cette option pour supprimer les observations comportant des données manquantes.

Onglet **Sorties**:

Résultats détaillés : activez cette option pour afficher les résultats détaillés des tests.

Tableau de synthèse : dans le cas où plusieurs échantillons ont été sélctionnés, activez cette option pour afficher un tableau de synthèse des p-values obtenues.

Onglet **Graphiques** :

Diagramme de dominance : activez cette option pour afficher un diagramme de dominance afin de comparer visuellement les échantillons.

Résultats

Les résultats affichés par XLSTAT correspondent aux différentes statistiques des tests sélectionnés, et à l'interprétation qui en découle.

Exemple

Un exemple de test de Wilcoxon signé est disponible sur le Centre d'aide XLSTAT :

http://www.xlstat.com/demo-wilcoxonf.htm

Un exemple de test de Wilcoxon signé est disponible sur le Centre d'aide XLSTAT :

http://www.xlstat.com/demo-mannwhitneyf.htm

Bibliographie

Cheung Y.K. Klotz J.H. (1997). The Mann Whitney Wilcoxon distribution using linked lists. Statistica Sinica, **7**, 805-813.

Hollander M. and Wolfe D. A. (1999). Nonparametric Statistical Methods, Second Edition. John Wiley and Sons, New York.

Lehmann E.L (1975). Nonparametrics: Statistical Methods Based on Ranks. Holden-Day, San Francisco.

Siegel S. and Castellan N. J. (1988). Nonparametric Statistics for the Behavioral Sciences, Second Edition. McGraw-Hill, New York.

Wilcoxon F. (1945). Individual comparisons by ranking methods. *Biometrics*, **1**, 80-83.

Comparaison de k échantillons (Kruskal-Wallis, Friedman, ...)

Utilisez cet outil pour comparer k échantillons indépendants (test de Kruskal- Wallis) ou appariés (test de Friedman, accéléré par GPU). Si une différence apparaît des procédures de comparaisons multiples sont à votre disposition pour effectuer les comparaisons.

Dans cette section :

Description

Afin de s'affranchir de l'hypothèse de normalité des échantillons nécessaire pour l'utilisation des tests de comparaisons multiples (proposés dans XLSTAT à la suite d'une ANOVA), des tests non paramétriques ont été proposés.

Comme pour les tests paramétriques, on distingue le cas où les échantillons sont indépendants (par exemple, dans le cas d'une comparaison des rendements de champs ayant des caractéristiques similaires mais traités avec trois types d'engrais différents), du cas où ils sont appariés (par exemple, dans le cas d'une comparaison des notations attribuées par 10 juges à 3 produits différents).

Comparaison de k échantillons indépendants

Le **test de Kruskal-Wallis** est souvent utilisé comme une alternative à l'ANOVA dans le cas où l'hypothèse de normalité n'est pas acceptable. Il permet de tester si k échantillons ($k \geq 2$) proviennent de la même population, ou de populations ayant des caractéristiques identiques, au sens d'un paramètre de position (le paramètre de position est conceptuellement proche de la médiane, mais le test de Kruskal-Wallis prend en compte plus d'information que la position au seul sens de la médiane).

Si on désigne par M_i le paramètre de position l'échantillon i, les hypothèses nulle H0 et alternative Ha du test de Kruskal-Wallis sont les suivantes :

- H0 : $M_1 = M_2 = \ldots = M_k$
- Ha : il existe au moins un couple (i, j) tel que $M_i \neq M_j$

Le calcul de la statistique K du test de Kruskal-Wallis fait intervenir, comme pour le test de Mann-Whitney, le rang des observations, une fois les k échantillons (ou groupes) mélangés. K est défini par :

$$K = \frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(N+1)$$

où n_i est la taille de l'échantillon i, N la somme des n_i, et R_i la somme des rangs pour l'échantillon i parmi l'ensemble des échantillons.

Lorsque $k = 2$ le test de Kruskal-Wallis est équivalent au test de Mann-Whitney, et la statistique K est équivalente à la statistique W.

Lorsqu'il y a des ex aequo, on utilise les rangs moyens pour les observations correspondantes, comme dans le cas du test de Mann-Whitney. La statistique K est alors définie par :

$$K = \frac{\frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(N+1)}{1 - \sum_{i=1}^{nd} (d_i^3 - d_i)/(N^3 - N)}$$

où nd est le nombre de valeurs distinctes, et d_i l'effectif correspondant à chacune de ces valeurs.

La statistique K est asymptotiquement distribuée suivant une loi du Khi² à $(k-1)$ degrés de liberté.

Comparaison de k échantillons appariés

Le **test de Friedman** est une alternative non paramétrique à l'ANOVA à deux facteurs sur échantillons appariés dans le cas où l'hypothèse de normalité n'est pas acceptable. Il permet de tester si k échantillons appariés ($k \geq 2$) de taille n, proviennent de la même population, ou de populations ayant des caractéristiques identiques, au sens d'un paramètre de position. Le contexte étant souvent celui de l'ANOVA à deux facteurs, on parle parfois de test de Friedman à k traitements et n blocs.

Si on désigne par M_i le paramètre de position de l'échantillon i, les hypothèses nulle H0 et alternative Ha du test de Friedman sont les suivantes :

- H0 : $M_1 = M_2 = \ldots = M_k$
- Ha : il existe au moins un couple (i, j) tel que $M_i \neq M_j$

Soit n la taille des k échantillons appariés. La statistique Q du test de Friedman est donnée par :

$$Q = \frac{12}{nk(k+1)} \sum_{i=1}^{k} R_i^2 - 3n(k+1)$$

où R_i est la somme des rangs pour l'échantillon i.

Lorsqu'il y a des ex aequo, on utilise les rangs moyens pour les observations correspondantes. La statistique Q est alors définie par :

$$Q = \frac{\frac{12}{nk(k+1)}\sum_{i=1}^{k} R_i^2 - 3n(k+1)}{1 - \sum_{j=1}^{n}\sum_{i=1}^{nd(j)}(d_{ij}^3 - d_{ij})/n/(k^3-k)}$$

où $nd(j)$ est le nombre de valeurs distinctes pour le block j, et d_{ij} l'effectif correspondant à chacune de ces valeurs.

Comme pour le test de Kruskal-Wallis, la p-value associée à une valeur donnée de Q peut être approximée par une loi du Khi² à $k-1$ degrés de liberté. Cette approximation est fiable lorsque $k \times n$ est plus grand que 30, la qualité dépendant aussi du nombre d'ex aequo. Les p-values associées à Q ont été tabulées pour le cas où (k = 3, n = 15) et (k = 4, n = 8) (Lehmann 1975, Hollander et Wolfe 1999).

Calcul de la p-value

Pour le calcul de la p-value associée à une valeur donnée de la statistique calculée, XLSTAT propose trois alternatives :

- Méthode asymptotique : la p-value est obtenue grâce à une approximation asymptotique de distribution de la statistique calculée.
- Méthode exacte : le calcul de la p-value exacte repose sur la distribution réelle de la statistique calculée. Ce calcul est très intensif numériquement.
- Méthode Monte Carlo : ce calcul est basé sur un rééchantillonnage aléatoire. L'utilisateur doit choisir le nombre de simulations (ou rééchantillonnages) à réaliser. Un intervalle de confiance autour de la p-value obtenue est fourni. Cet intervalle sera bien entendu d'autant plus resserré que le nombre de simulations est important.

Afin d'éviter un blocage d'Excel dans le cas des deux dernières méthodes, XLSTAT donne la possibilité à l'utilisateur de fixer le temps maximum, exprimé en secondes, qu'il souhaite consacrer à la recherche de la p-value.

Comparaisons multiples par paires

Que ce soit pour le test de Kruskal-Wallis, ou le test de Friedman, si la p-value est telle que l'on doit rejeter l'hypothèse H0, alors au moins un échantillon (ou groupe) est différent d'un autre. Afin d'identifier quels échantillons sont responsables du rejet de H0, il est possible d'utiliser une procédure de **comparaisons multiples**.

Pour le test de Kruskal-Wallis trois méthodes de comparaisons multiples sont proposées :

- Dunn (1963) : propose une méthode basée sur la comparaison des moyennes des rangs, ces derniers étant ceux utilisés pour le calcul du K, en utilisant une distribution normale asymptotique pour la différence standardisée de la moyenne des rangs.
- Conover et Iman (1999) : proche de la méthode de Dunn, cette méthode utilise une distribution de Student. Elle correspond à un test de Student réalisé sur les rangs.
- Steel-Dwass-Critchlow-Fligner (1984) : cette méthode, plus complexe mais recommandée par Hollander (1999), nécessite le recalcul des rangs pour chaque combinaison deux à deux des échantillons. La statistique Wij est calculée pour chaque combinaison. XLSTAT calcule ensuite la p-value correspondante en utilisant la distribution asymptotique de la statistique.

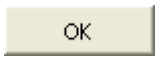
Pour le test de Friedman, la méthode de comparaisons multiples proposée est celle de Nemenyi (1963). Cette méthode est proche de celle de Dunn, mais prend en compte l'appariement des données.

Pour les méthodes de Dunn, de Conover et Iman, afin de prendre en compte le fait qu'il y a k(k-1)/2 comparaisons réalisées, la correction du niveau de signification proposée par Bonferroni peut être appliquée. Le niveau de signification utilisé pour les comparaisons deux à deux est alors :

$$\alpha' = \frac{2\alpha}{k(k-1)}$$

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.

Onglet **Général** :

Données : si le format de données sélectionné est « une colonne par variable », sélectionnez les données correspondant aux différents échantillons sur la feuille Excel. Si le format de données sélectionné est « une colonne par échantillon » ou « échantillons appariés », sélectionnez les colonnes de données correspondant aux différents échantillons.

Identifiant d'échantillon : si le format de données sélectionné est « une colonne par variable », sélectionnez les données identifiant les k échantillons auxquels les données sélectionnées correspondent.

Format des données : choisissez le format des données.

- **Une colonne/ligne par échantillon** : activez cette option pour sélectionner une colonne (ou ligne en mode lignes) par échantillon.
- **Une colonne/ligne par variable** : activez cette option pour que XLSTAT fasse autant de tests qu'il y a de colonnes/lignes, sachant que chaque colonne/ligne doit contenir le même nombre de lignes/colonnes, et qu'un identifiant d'échantillon permettant d'affecter chaque observation à un échantillon doit par ailleurs être sélectionné.
- **Echantillons appariés** : activez cette option pour faire des tests sur échantillons appariés. Vous devez alors sélectionner une colonne (ou ligne en mode lignes) par échantillon, tout en veillant à ce que les échantillons soient de même taille.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des colonnes/lignes : activez cette option si la première ligne (mode colonnes) ou colonne (mode lignes) des données sélectionnées contient des libellés.

Test de Kruskal-Wallis : activez cette option pour utiliser le test de Kruskal-Wallis (voir description).

Test de Friedman : activez cette option pour utiliser le test de Friedman (voir description).

Comparaisons multiples par paires : activez cette option pour calculer les tests de comparaisons multiples par paires (voir description). Dans le cas du test de Kruskal-Wallis, trois méthodes sont proposées.

- **Correction de Bonferroni** : activez cette option pour utiliser le niveau de signification corrigé de Bonferroni.

Onglet **Options** :

Niveau de signification (%) : entrez la valeur du niveau de signification pour les tests (valeur par défaut : 5%).

En fonction du test utilisé, plusieurs méthodes de calcul de la p-value sont proposées. Choisissez entre la méthode asymptotique, exacte ou Monte Carlo (voir description). Dans le cas des méthodes exacte et Monte Carlo, vous pouvez préciser quel temps maximum vous souhaitez consacrer à la recherche de la p-value.

Onglet **Données manquantes** :

Ne pas accepter les valeurs manquantes : activez cette option pour que XLSTAT empêche la poursuite des calculs si des valeurs manquantes sont détectées.

Supprimer les observations : activez cette option pour supprimer les observations comportant des données manquantes.

Onglet **Sorties** :

Statistiques descriptives : activez cette option pour afficher les statistiques descriptives des échantillons.

Résultats

Les résultats affichés par XLSTAT correspondent aux différentes statistiques des tests sélectionnés, et à l'interprétation qui en découle.

Exemple

Un exemple de test de Kruskal-Wallis est disponible sur le Centre d'aide XLSTAT :

http://www.xlstat.com/demo-kruskalf.htm

Un exemple de test de Friedman est disponible sur le Centre d'aide XLSTAT :

http://www.xlstat.com/demo-friedmanf.htm

Bibliographie

Conover W.J. (1999). Practical Nonparametric Statistics, 3rd edition, Wiley.

Critchlow D.E. (1980). Metric Methods for Analyzing Partially Ranked Data. Lecture Notes in Statistics 34, Springer-Verlag.

Dunn O.J. (1964). Multiple Comparisons Using Rank Sums. *Technometrics*, **6(3)**, 241-252.

Dwass M. (1960). Some k-sample rank-order tests. In: I. Olkin, S. G. Ghurye, W. Hoeffding, W. G. Madow & H. B. Mann, editors. Contributions to probability and statistics. Essays in honor of Harold Hotelling. Stanford University Press, 198-202.

Fligner M. A. (1984). A note on two-sided distribution-free treatment versus control multiple comparisons. *Journal. Am. Statist. Assoc.*, **79**, 208-211.

Friedman M. A. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *J. Amer. Statist. Assoc*., **32**, 675-701.

Friedman M. A. (1940). A comparison of alternative tests of significance for the test of m rankings. *Annals of Mathematical Statistics*, **11**, 86-92.

Hollander M. and Wolfe D. A. (1999). Nonparametric Statistical Methods, Second Edition. John Wiley and Sons, New York.

Lehmann E.L (1975). Nonparametrics: Statistical Methods Based on Ranks. Holden-Day, San Francisco.

Nemenyi P. (1963). Distribution-Free Multiple Comparisons. Unpublished Ph.D Thesis.

Siegel S. and Castellan N. J. (1988). Nonparametric Statistics for the Behavioral Sciences, Second Edition. McGraw-Hill, New York.

Steel R. G. D. (1961). Some rank sum multiple comparison tests. *Biometrics*, **17**, 539-552.

Tests de Durbin-Skillings-Mack

Utilisez cet outil pour tester si k traitements évalués suivant un plan en blocs incomplets (équilibrés ou non) sont identiques ou différents (accéléré par GPU).

Dans cette section :

Description

Le test proposé par Durbin en 1951 a pour but de permettre d'analyser avec rigueur et en utilisant une méthode non paramétrique (ne faisant donc pas d'hypothèse sur la distribution des mesures faites) les résultats d'une étude construite suivant un plan en blocs incomplets équilibrés. Skillings et Mack (1981) ont eux proposé une extension pour des plans en blocs plus généraux.

Plans en blocs

Un plan en blocs est un plan d'expériences dans lequel on étudie l'influence d'au moins deux facteurs sur un ou plusieurs phénomènes. On sait que l'un des facteurs a par construction un effet important, sans que l'on puisse agir dessus, mais ce n'est pas celui qui nous intéresse. On veut donc pouvoir s'assurer que ce facteur ne perturbera pas les analyses que l'on effectuera une fois les données collectées. Pour cela on fait en sorte que les différents niveaux des autres facteurs soient aussi bien représentés dans chacun des blocs (les modalités du facteur bloc).

Dans le cas d'une étude sur des produits évalués par des juges, nous avons un facteur bloc qui correspond aux juges, et un facteur que l'on souhaite particulièrement étudié, le facteur produit.

Un plan en blocs complets est un plan dans lequel tous les niveaux des facteurs étudiés sont présents une fois à l'intérieur de chaque bloc. Cela correspond, pour un plan sensoriel, au cas où tous les produits sont vus une fois par l'ensemble des juges.

Un plan en **blocs incomplets** est un plan dans lequel tous les niveaux des facteurs étudiés ne sont pas présents dans chaque bloc. Il est **équilibré** si chaque niveau de chaque facteur étudié est présent un même nombre r de fois et si chaque couple de niveaux de chaque facteur étudié est présent un même nombre de fois λ.

Si t est le nombre de traitements (les produits par exemple) étudiés, b le nombre de blocs (les juges par exemple), k le nombre de traitements présentés dans un bloc, on montre que les conditions suivantes sont nécessaires (mais non suffisantes) pour avoir un plan en blocs incomplets équilibrés:

$$bk = tr$$

et

$$r(k-1) = \lambda(t-1)$$

Tests de Durbin et Skillings- Mack

Les tests de Durbin et Skillings-Marck sont une alternative au test de Friedman (1937) qui correspond, lui, au cas du plan en blocs complets.

Les hypothèses nulle et alternative associées à ces tests sont comme pour le test de Friedman :

- H_0 : les t traitements ne sont pas significativement différents.
- H_a : au moins l'un des traitements est différent d'un autre.

La statistique proposée par Durbin est donnée par

$$Q = \frac{12(t-1)}{rt(k-1)(k+1)} \sum_{j=1}^{t} \left(R_j - \frac{r(k+1)}{2} \right)^2$$

où R_j est la somme sur les b blocs pour le traitement j des rangs R_{ij} pour le traitement j et le bloc i.

Dans le cas où des ex-æquo sont présents au sein de blocs, une correction doit être apportée. On a alors

$$Q = \frac{(t-1)}{A-C} \left(\left(\sum_{j=1}^{t} R_j^2 \right) - rC \right) \text{ avec } A = \sum i = 1^b \sum_{j=1}^{t} R_{ij}^2 \text{ et } C = \frac{bk(k+1)}{4}$$

Cette statistique a la propriété d'être asymptotiquement distribuée suivant une loi du χ^2 à $t-1$ degrés de liberté. Alvo et Cabilio (1995) proposent une statistique modifiée ayant, selon Conover (1999) de meilleures propriétés asymptotiques :

$$F = \frac{\frac{Q}{(t-1)}}{(b(k-1) - Q)(b(k-1) - t + 1)}$$

Cette statistique a la propriété d'être asymptotiquement distribuée suivant un F de Fisher à $t-1$ et $b(k-1)-t+1$ degrés de liberté.

Le calcul de la statistique T de Skillings et Mack qui permet de traiter les cas de plans en blocs incomplets non équilibrés est plus complexe. Les valeurs manquantes sont remplacées par une moyenne, et un poids compensatoire est appliqué aux blocs ayant des valeurs manquantes. La statistique T est distribuée suivant une loi du χ^2 à $t-1$ degrés de liberté.

Calcul de la p-value

Pour le calcul de la p-value associée aux différentes statistiques, XLSTAT propose deux alternatives :

- Méthode asymptotique : la p-value est obtenue grâce à une approximation de la loi de Q ou de F. Cette approximation est d'autant plus fiable que le nombre de blocs et/ou de traitements est important.
- Méthode Monte Carlo : ce calcul est basé sur un rééchantillonnage aléatoire. L'utilisateur doit choisir le nombre de simulations (ou rééchantillonnages) à réaliser. Un intervalle de confiance autour de la p-value obtenue est fourni. Cet intervalle sera bien entendu d'autant plus resserré que le nombre de simulations est important.

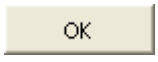
Afin d'éviter un blocage d'Excel dans le cas de la méthode Monte Carlo, XLSTAT donne la possibilité à l'utilisateur de fixer le temps maximum, exprimé en secondes, qu'il souhaite consacrer à la recherche de la p-value.

Comparaisons multiples par paires

Si la p-value est telle que l'on doit rejeter l'hypothèse H0, alors au moins un traitement est différent d'un autre. Afin d'identifier quel(s) traitement(s) est/sont responsable(s) du rejet de H_0, il est possible d'utiliser une procédure de comparaisons multiples. XLSTAT propose pour le test de Durbin celle décrite dans Conover (1999). Les tests de comparaisons multiples ne sont pas faits dans le cas des blocs incomplets non équilibrés.

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas (mode colonnes), XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite (mode lignes), XLSTAT considère que les variables sont en lignes et les observations en colonnes.

Onglet **Général** :

Tableau Individus/Traitements : sélectionnez un tableau dont les lignes correspondent aux individus (ou blocs) et les colonnes aux traitements (ou vice versa si le mode lignes est activé). Si les libellés des traitements ont été sélectionnés, veillez à ce que l'option « libellés des traitements » soit activée.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des traitements : activez cette option si des en-têtes de colonne ont été sélectionnés (ou des en-têtes de ligne en mode lignes).

Comparaisons multiples par paires : activez cette option pour calculer les tests de comparaisons multiples par paires.

Onglet **Options** :

Niveau de signification (%) : entrez le niveau de signification à utiliser pour les différents tests (valeur par défaut : 5%).

Pour le calcul de la p-value, vous pouvez choisir pour entre la méthode asymptotique ou la méthode de Monte Carlo (voir la section description). Dans le cas de la méthode Monte Carlo, vous pouvez préciser le nombre de rééchantillonnages, et quel temps maximum vous souhaitez consacrer à l'estimation de la p-value.

Onglet **Sorties** :

Statistiques descriptives : activez cette option pour calculer et afficher les statistiques descriptives pour les différents traitements.

Résultats

Statistiques descriptives : dans ce tableau sont affichées les statistiques descriptives correspondant aux différents traitements.

Les résultats qui correspondent au test de Durbin (cas d'un plan en blocs incomplets équilibrés) ou de Skillings-Mack dans un cas plus général sont ensuite affichés. Une interprétation du test est fournie, suivie des comparaisons multiples afin d'identifier les traitements responsables du rejet de l'hypothèse nulle si elle a été rejetée.

Exemple

Un exemple de test de Durbin est disponible sur le Centre d'aide XLSTAT :

http://www.xlstat.com/demo-durbinf.htm

Bibliographie

Alvo M. and Cabilio P. (1995). Approximate and exact distributions of rank tests for balanced incomplete block designs. Communications in Statistics - Theory and Methods, 24(12), 3073-3121.

Conover W.J. (1999). Practical Nonparametric Statistics, 3rd edition, Wiley.

Durbin J. (1951). Incomplete blocks in ranking experiments. *Brit. J. Statist. Psych.*, **4**, 85-90.

Friedman M. A. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *J. Amer. Statist. Assoc*., **32**, 675-701.

Hollander M. and Wolfe D. A. (1999). Nonparametric Statistical Methods, Second Edition. John Wiley and Sons, New York.

Siegel S. and Castellan N. J. (1988). Nonparametric Statistics for the Behavioral Sciences, Second Edition. McGraw-Hill, New York.

Skillings J. H. and Mack G. A. (1981). On the use of a Friedman-type statistic in balanced and unbalanced block designs. *Technometrics*, **23**, 171-177.

Test de Page

Utilisez cet outil pour tester si k traitements évalués suivant un plan en b blocs incomplets (équilibrés ou non) sont identiques ou si un classement supposé peut être accepté (accéléré par GPU).

Dans cette section :

Description

Ce test a d'abord été proposé par Page en 1963 pour des plans en blocs complets, puis il a été étendu à des plans en blocs incomplets par Alvo et Cabilio (2005). Ce test est une méthode non paramétrique (ne faisant donc pas d'hypothèse sur la distribution des mesures faites) qui permet d'analyser les résultats d'une étude dont le but est de mettre en évidence l'absence d'effets de traitements (le terme vient du domaine médical, mais en marketing, cela peut être un produit ou une offre), ou alternativement l'existence d'un classement pressenti des traitements. Ce test diffère du test de Friedman ou de ses variantes pour des blocs incomplets du fait de l'hypothèse alternative qui suppose ici un ordre.

Plans en blocs

Un plan en blocs est un plan d'expériences dans lequel on étudie l'influence d'au moins deux facteurs sur un ou plusieurs phénomènes. On sait que l'un des facteurs a par construction un effet important, sans que l'on puisse agir dessus, mais ce n'est pas celui qui nous intéresse. On veut donc pouvoir s'assurer que ce facteur ne perturbera pas les analyses que l'on effectuera une fois les données collectées. Pour cela on fait en sorte que les différents niveaux des autres facteurs soient aussi bien représentés dans chacun des blocs (les modalités du facteur bloc).

Dans le cas d'une étude sur des produits évalués par des juges, nous avons un facteur bloc qui correspond aux juges, et un facteur que l'on souhaite particulièrement étudié, le facteur produit.

Un plan en blocs complets est un plan dans lequel tous les niveaux des facteurs étudiés sont présents une fois à l'intérieur de chaque bloc. Cela correspond, pour un plan sensoriel, au cas où tous les produits sont vus une fois par l'ensemble des juges.

Un plan en **blocs incomplets** est un plan dans lequel tous les niveaux des facteurs étudiés ne sont pas présents dans chaque bloc. Il est **équilibré** si chaque niveau de chaque facteur étudié

est présent un même nombre r de fois et si chaque couple de niveaux de chaque facteur étudié est présent un même nombre de fois l.

Si t est le nombre de traitements (les produits par exemple) étudiés, b le nombre de blocs (les juges par exemple), k le nombre de traitements présentés dans un bloc, on montre que les conditions suivantes sont nécessaires (mais non suffisantes) pour avoir un plan en blocs incomplets équilibrés:

$$bk = tr$$

et

$$r(k-1) = l(t-1)$$

Test de Page

Si $T_1, T_2, \ldots, T_t$ désigne les t traitements, les hypothèses nulles et alternatives associées à ce test sont :

- H0 : les t traitements ne sont pas différents.
- Ha : $T_1 \geq T_2 \geq \ldots \geq T_t$

ou

- Ha : $T_1 \leq T_2 \leq \ldots \leq T_t$

Avec pour les hypothèses alternatives au moins une inégalité stricte.

La statistique proposée par Page est donnée par

$$L = \sum_{j=1}^{t} (jR_j)$$

Page a tabulé cette statistique, mais il donne également une statistique asymptotique qui suit une loi du Khi² à 1 degré de liberté. Conover utilise la statistique suivante qui suit une loi normale standard :

$$z = \frac{12L - 3bt(t+1)^2}{\sqrt{bt^2(t^2-1)(t+1)}}$$

où b est le nombre de blocs et t le nombre de traitements. Dans le cas où des ex-aequo sont présents le terme de variance est modifié.

Dans le cas des blocs incomplets, Alvo et Cabilio (2005) proposent une statistique alternative dont la valeur est identique dans le cas complet et qui a les mêmes propriétés asymptotiques.

Calcul de la p-value

Pour le calcul de la p-value associée aux différentes statistiques, XLSTAT propose deux alternatives :

- Méthode asymptotique : la p-value est obtenue grâce à l'approximation par la statistique qui suit une loi normale standard. Cette approximation est d'autant plus fiable que le nombre de blocs et/ou de traitements est important.
- Méthode Monte Carlo : ce calcul est basé sur un rééchantillonnage aléatoire. L'utilisateur doit choisir le nombre de simulations (ou rééchantillonnages) à réaliser. Un intervalle de confiance autour de la p-value obtenue est fourni. Cet intervalle sera bien entendu d'autant plus resserré que le nombre de simulations est important.

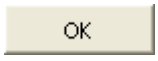
Afin d'éviter un blocage d'Excel dans le cas de la méthode Monte Carlo, XLSTAT donne la possibilité à l'utilisateur de fixer le temps maximum, exprimé en secondes, qu'il souhaite consacrer à la recherche de la p-value.

Comparaisons multiples par paires

Si la p-value est telle que l'on doit rejeter l'hypothèse H0, alors au moins un traitement est supérieur ou inférieur à un moins un autre. Afin d'identifier quel(s) traitement(s) est/sont responsable(s) du rejet de H0, il est possible d'utiliser la procédure de comparaisons multiples proposée par Cabilio et Peng (2008), soit avec un calcul de p-value basé sur une approximation asymptotique par une loi normale, soit un calcul basé sur des rééchantillonnages Monte Carlo.

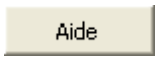
Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.



: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas (mode colonnes), XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite (mode lignes), XLSTAT considère que les variables sont en lignes et les observations en colonnes.

Onglet **Général**:

Tableau Individus/Traitements : sélectionnez un tableau dont les lignes correspondent aux individus (ou blocs) et les colonnes aux traitements (ou vice versa si le mode lignes est activé). Si les libellés des traitements ont été sélectionnés, veillez à ce que l'option « libellés des traitements » soit activée.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des traitements : activez cette option si des en-têtes de colonne ont été sélectionnés (ou des en-têtes de ligne en mode lignes).

Comparaisons multiples par paires : activez cette option pour calculer les tests de comparaisons multiples par paires. Vous avez le choix entre la procédure décrite par Cabilio et Peng utilisant une p-value asymptotique ou la même procédure avec des p-values calculées avec des rééchantillonnages Monte Carlo (voir la section description).

Onglet **Options**:

Niveau de signification (%) : entrez le niveau de signification à utiliser pour les différents tests (valeur par défaut : 5%).

Pour le calcul de la p-value, vous pouvez choisir pour entre la méthode asymptotique ou la méthode de Monte Carlo (voir la section description). Dans le cas de la méthode Monte Carlo, vous pouvez préciser le nombre de rééchantillonnages, et quel temps maximum vous souhaitez consacrer à l'estimation de la p-value.

Onglet **Sorties**:

Statistiques descriptives : activez cette option pour calculer et afficher les statistiques descriptives pour les différents traitements.

Résultats

Statistiques descriptives : dans ce tableau sont affichées les statistiques descriptives correspondant aux différents traitements.

Les résultats qui correspondent au test de Page (cas d'un plan complet) ou de Alvo et Cabilio dans un cas plus général sont ensuite affichés. Une interprétation du test est fournie, suivie des comparaisons multiples afin d'identifier les traitements responsables du rejet de l'hypothèse nulle si elle a été rejetée.

Exemple

Un exemple de test de Page est disponible sur le Centre d'aide XLSTAT :

http://www.xlstat.com/demo-pagef.htm

Bibliographie

Alvo M. and Cabilio P. (1995). Testing ordered alternatives in the presence of incomplete data. *Journal of the American Statistical Association* , **90(431)**, 1015-1024.

Cabilio P. and Peng J. (2008). Multiple rank-based testing for ordered alternatives with incomplete data. *Statistics and Probability Letters*, **78**, 2609-2613.

Conover W.J. (1999). Practical Nonparametric Statistics, 3rd edition, Wiley.

Page E. B. (1963). Ordered hypotheses for multiple treatments: A significance test for linear ranks". *Journal of the American Statistical Association*, **58(301)**, 216-230.

Test Q de Cochran

Utilisez cet outil pour comparer $k \geq 2$ échantillons appariés dont les valeurs sont binaires (accéléré par GPU).

Dans cette section :

Description

Le test Q de Cochran est présenté sous deux angles différents. Certains auteurs le présentent comme un cas particulier du test de Friedman (comparaison de k échantillons appariés) pour le cas où la variable mesurée est binaire (Lehmann, 1975), d'autres le présentent comme un test d'homogénéité marginale pour un tableau de contingence à k dimensions (Agresti, 1990).

Les hypothèses nulles et alternatives associées au test Q de Cochran sont alors soit,

- H0 : les k traitements ne sont pas différents.
- Ha : au moins l'un des traitements est différent d'un autre.

soit,

- H0 : les k distributions sont marginalement homogènes.
- Ha : les k distributions sont marginalement inhomogènes.

XLSTAT utilise la première représentation, plus classique, et utilise la terminologie commune de « traitements » pour les k échantillons comparés.

Deux formats sont proposés pour la saisie des données :

- vous pouvez sélectionner des données sous un format brut, correspondant à la saisie progressive des résultats. Chaque colonne correspond à un traitement et chaque ligne correspond à un individu.
- vous pouvez aussi sélectionner des données sous un format « groupé ». Chaque colonne correspond à un traitement, et chaque ligne à une combinaison de réponses possibles

pour les k traitements. Vous devez ensuite saisir les effectifs correspondant à chacune des combinaisons (champ « Effectifs » dans la boîte de dialogue).

Calcul de la p-value

Pour le calcul de la p-value associée à une valeur donnée de la statistique calculée, XLSTAT propose en fonction des tests trois alternatives :

- Méthode asymptotique : la p-value est obtenue grâce à une approximation asymptotique de la distribution de la statistique calculée.
- Méthode exacte : le calcul de la p-value exacte repose sur la distribution réelle de la statistique calculée. Ce calcul est parfois très intensif numériquement.
- Méthode Monte Carlo : ce calcul est basé sur un rééchantillonnage aléatoire. L'utilisateur doit choisir le nombre de simulations (ou rééchantillonnages) à réaliser. Un intervalle de confiance autour de la p-value obtenue est fourni. Cet intervalle sera bien entendu d'autant plus resserré que le nombre de simulations est important.

Comparaisons multiples par paires

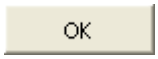
Lorsque l'hypothèse H0 du test de Cochran est rejetée et que l'on conclue qu'au moins un des traitements est différent d'un autre. Il peut être intéressant de comparer les traitements deux à deux. Pour cela XLSTAT propose deux méthodes différentes :

- La méthode **Différence critique (Sheskin)** expliquée dans Sheskin (2011) et initialement développée par Marascuilo et McSweeney (1977). Cette méthode calcule une valeur critique de différence CD. Lorsque la différence entre deux traitements est supérieure à CD, on conclue qu'il y a une différence significative entre les deux traitements.
- La procédure de **McNemar(Bonferroni)** consiste à réaliser des tests de McNemar entre les différentes paires de traitements en appliquant la correction de Bonferroni. Cette correction consiste à diviser le niveau de significativité α par le nombre total de comparaisons $(k \times (k-1)/2)$. Les p-values des tests de McNemar sont obtenues grâce à une approximation asymptotique de la distribution de la statistique calculée. Une correction de continuité est également apportée. Vous pouvez vous reporter à la section drelative au test de McNemar pour plus de détails.

La méthode **Différence critique (Sheskin)** est l'option préconisée car la valeur critique est calculée sur l'ensemble des traitements contrairement aux tests multiples de McNemar qui n'utilisent que les données correspondant à la paire de traitements à comparer.

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas (mode colonnes), XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite (mode lignes), XLSTAT considère que les variables sont en lignes et les observations en colonnes.

Onglet **Général**:

Tableau Individus/Traitements : sélectionnez un tableau dont les lignes correspondent aux individus (ou blocs) et les colonnes aux traitements dans le cas du « mode colonnes », ou vice-versa dans le « mode lignes ». Si les libellés des traitements ont été sélectionnés, veillez à ce que l'option « libellés des traitements » soit activée.

Format des données :

- Tableau Individus/Traitements :
- **Brut** : choisissez cette option si les données sont brutes, par opposition à groupées.
- **Groupé** : choisissez cette option si les données correspondent à un regroupement ou si elles sont pondérées. Vous devez alors choisir les poids associés à chacune des lignes du tableau sélectionné.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des traitements : activez cette option si des en-têtes de colonne ont été sélectionnés.

Poids : sélectionnez les poids associés aux individus. Les poids être impérativement supérieurs à 0. Si un en-tête de colonne a été sélectionné, veuillez vérifier que l'option « Libellés des

traitements » est activée.

Comparaisons multiples par paires : activez cette option pour calculer les tests de comparaisons multiples par paires. Deux méthodes sont proposées :

- **Différence critique (Sheskin)** : cette procédure est basée sur une valeur critique globale.
- **McNemar(Bonferroni)** : pour réaliser des tests de McNemar avec un seuil modifié.

Onglet **Options**:

Niveau de signification (%) : entrez le niveau de signification à utiliser pour les différents tests (valeur par défaut : 5%).

En fonction du test utilisé, plusieurs méthodes de calcul de la p-value sont proposées. Choisissez entre la méthode asymptotique, exacte ou Monte Carlo (voir la section description). Dans le cas des méthodes exacte et Monte Carlo, vous pouvez préciser quel temps maximum vous souhaitez consacrer à la recherche de la p-value.

Onglet **Sorties**:

Statistiques descriptives : activez cette option pour calculer et afficher les statistiques descriptives pour les différents traitements.

Résultats

Statistiques descriptives : dans ce tableau sont affichées les statistiques descriptives correspondant aux différents traitements.

Les résultats qui correspondent au test de Cochran Q sont ensuite affichés. Une interprétation du test est fournie, suivie des comparaisons mutliples afin d'identifier les traitements responsables du rejet de l'hypothèse nulle si elle a été rejetée.

Exemple

Un exemple de test de Cochran est disponible sur le Centre d'aide XLSTAT :

http://www.xlstat.com/demo-cochranqf.htm

Bibliographie

Agresti A. (1990). Categorical Data Analysis. John Wiley and Sons, New York.

Cochran W.G. (1950). The comparison of percentages in matched samples. *Biometrika*, **37**, 256-266.

Lehmann E.L (1975). Nonparametrics: Statistical Methods Based on Ranks. Holden-Day, San Francisco.

Marascuilo L.A. and McSweeney M. (1977). Nonparametric and Distribution- Free Methods for the Social Sciences. Brooks/Cole, Monterey, CA.

Sheskin, D.J. (2011). Handbook of Parametric and Non-Parametric Statistical Procedures. 5th Edition, Chapman & Hall/CRC, London.

Test de McNemar

Utilisez cet outil pour comparer 2 échantillons appariés dont les valeurs sont binaires, et éventuellement synthétisées dans un tableau de contingence 2x2.

Dans cette section :

Description

Le test de McNemar est équivalent au test du Q de Cochran dans le cas où l'on a que deux traitements. Comme pour le test de Cochran, la variable étudiée est binaire. Le test de McNemar présente néanmoins deux avantages :

- Le calcul de la p-value exacte est possible (Lehmann, 1975) ;

- Les données peuvent être présentées sous la forme d'un tableau de contingence à deux dimensions.

Dans le cas d'un test bilatéral, les hypothèses nulle (H0) et alternative (Ha) sont les suivantes :

- H0 : Traitement 1 = Traitement 2
- Ha : Traitement 1 $\neq$ Traitement 2

Dans le cas unilatéral, il faut distinguer le test unilatéral à gauche (ou inférieur) et le test unilatéral à droite (ou supérieur). Dans le test unilatéral à gauche, les hypothèses sont les suivantes :

- H0 : Traitement 1 = Traitement 2
- Ha : Traitement 1 < Traitement 2

Pour le test unilatéral à droite, les hypothèses sont les suivantes :

- H0 : Traitement 1 = Traitement 2
- Ha : Traitement 1 > Traitement 2

Trois formats sont proposés pour la saisie des données :

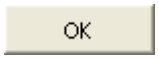
- vous pouvez sélectionner des données sous un format brut, correspondant à la saisie progressive des résultats. Chaque colonne correspond à un traitement et chaque ligne correspond à un individu.

- vous pouvez aussi sélectionner des données sous un format « groupé ». Chaque colonne correspond à un traitement, et chaque ligne à une combinaison de réponses possibles pour les deux traitements (il y a quatre combinaisons possibles). Vous devez ensuite saisir les effectifs correspondant à chacune des combinaisons (champ « effectifs » dans la boîte de dialogue).

- vous pouvez aussi sélectionner un tableau de contingence à deux lignes et deux colonnes. Dans le cas où ce format est choisi, les traitements 1 et 2 sont considérés comme correspondant respectivement aux lignes et aux colonnes. Les cas de succès des traitements (ou réponse positive) sont considérés comme correspondant à la première ligne (traitement 1) ou à la première colonne (traitement 2) du tableau de contingence.

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas (mode colonnes), XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite (mode lignes), XLSTAT considère que les variables sont en lignes et les observations en colonnes.

Onglet **Général**:

Tableau Individus/Traitements / Tableau de contingence : dans le cas d'un tableau « Individus/Traitements », sélectionnez un tableau dont les lignes correspondent aux individus (ou blocs) et les colonnes aux traitements dans le cas du « mode colonnes », ou vice-versa dans le « mode lignes ». Dans le cas d'un tableau de contingence, sélectionnez les données du

tableau de contingence. Si les libellés des traitements ont été sélectionnés, veillez à ce que l'option « libellés des traitements » ou « libellés inclus » soit activée.

Format des données :

- **Tableau Individus/Traitements** : choisissez cette option si vos données correspondent à un tableau individus/traitements
- **Brut** : choisissez cette option si les données sont brutes, par opposition à groupées.
- **Groupé** : choisissez cette option si les données correspondent à un regroupement ou si elles sont pondérées. Vous devez alors choisir les poids associés à chacune des lignes du tableau sélectionné.
- **Tableau de contingence** : activez cette option si vos données sont contenues dans un tableau de contingence.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des traitements/Libellés inclus : activez cette option si des en-têtes ont été sélectionnés. Dans le cas d'un tableau de contingence, les libellés des lignes et des colonnes doivent être sélectionnés.

Poids : sélectionnez les poids associés aux individus. Les poids doivent être impérativement supérieurs à 0. Si un en-tête de colonne a été sélectionné, veuillez vérifier que l'option « Libellés des traitements » est activée.

Code (réponse positive) : entrez le code qui correspond dans vos données à une réponse positive (ou à un succès).

Onglet **Options**:

Niveau de signification (%) : entrez le niveau de signification à utiliser pour les différents tests (valeur par défaut : 5%).

p-value exacte : activez cette option pour calculer la p-value exacte.

Correction de continuité : activez cette option si vous souhaitez que XLSTAT utilise la correction de continuité si le calcul d'une p-value exacte n'est pas demandé.

Onglet **Sorties**:

Cet onglet n'est visible que lorsque le format choisi est « Tableau individus/traitements ».

Statistiques descriptives : activez cette option pour calculer et afficher les statistiques descriptives pour les différents traitements.

Tableau de contingence : activez cette option pour afficher le tableau de contingence 2x2.

Résultats

Statistiques descriptives : dans ce tableau sont affichées les statistiques descriptives correspondant aux deux traitements.

Tableau de contingence : le tableau de contingence 2x2 est affiché.

Les résultats qui correspondent au test de McNemar sont ensuite affichés. Une interprétation du test est fournie.

Exemple

Un exemple de test de McNemar est disponible sur le Centre d'aide XLSTAT :

http://www.xlstat.com/demo-mcnemarf.htm

Bibliographie

Agresti A. (1990). Categorical Data Analysis. John Wiley and Sons, New York.

McNemar Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, **12**, 153-157.

Lehmann E.L (1975). Nonparametrics: Statistical Methods Based on Ranks. Holden-Day, San Francisco.

Test de Cochran-Mantel-Haenszel

Utilisez cet outil pour tester l'hypothèse d'indépendance sur une série de tableaux de contingence correspondant à une expérience croisant deux variables catégorielles, avec une variable de contrôle prenant plusieurs valeurs.

Dans cette section :

Description

Imaginons le cas d'un laboratoire travaillant sur un nouvel antifongique. Dans le but de définir la dose et la forme galénique adéquate, une expérience est menée avec quatre niveaux de dose et sous deux conditionnements différents (pommade ou gel douche). Pour chaque niveau de dose, le test est effectué sur une vingtaine de patients, équitablement répartis pour chaque conditionnement. Pour chaque patient, on évalue si le traitement est efficace ou non. Les résultats se présentent donc sous la forme d'un tableau de contingence à trois dimensions, ou plus simplement sous forme de 4 tableaux de contingence à deux dimensions. La variable correspondant à la dose est appelée variable de contrôle.

On pourrait vouloir faire un test d'indépendance sur le tableau résultant de la somme des 4 tableaux de contingence, néanmoins on risquerait dans ce cas de conclure à l'indépendance pour la seule raison que le sous-tableau ayant l'effectif le plus important correspond à un cas d'indépendance, alors que pour les autres tableaux on a une dépendance, ou parce que des dépendances diverses sont annulées par la somme.

Cochran (1954) puis Mantel et Haenszel (1959) ont proposé un test permettant de tester s'il y a indépendance entre les deux lignes et les deux colonnes des tableaux de contingence, sachant que les tableaux sont indépendants entre eux (pour chaque dose il s'agit de patients différents), et en conditionnant par rapport aux sommes marginales de chacun, comme dans le test standard d'indépendance sur tableaux de contingence.

Le test communément appelé test de Cochran-Mantel-Haenszel (CMH) s'appuie sur la statistique M^2 définie par :

$$M^2 = \frac{\left(\left| \sum_{i=1}^{k} (n_{11i} - n_{1+i} n_{+1i} / n_{++i} \right| - \frac{1}{2} \right)^2}{\sum_{i=1}^{k} n_{1+i} n_{2+i} n_{+1i} n_{+2i} / (n_{++i}^2 (n_{++i}^2 - 1))}$$

Cette statistique suit asymptotiquement une loi du χ^2 à 1 degré de liberté. Connaissant M^2, on peut donc connaître la p-value, et connaissant le risque de première espèce, α, on peut déterminer la valeur critique. Il est aussi possible comme pour le test d'indépendance sur un tableau de contingence de calculer la p-value exacte dans le cas où les tableaux de contingence sont de taille 2x2. L'utilisation de la valeur absolue et la soustraction de $-\frac{1}{2}$ ainsi que la division par $\left(n_{++i}^2 - 1\right)$ au lieu de n_{++i}^2 correspondent à une correction de continuité proposée par Mantel et Haenszel. Son utilisation est vivement conseillée. XLSTAT permet néanmoins de la désactiver.

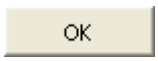
On peut noter au numérateur que l'on mesure l'écart à l'indépendance pour chaque case en haut à gauche du tableau de contingence et que l'on fait ensuite la somme des écarts. Si les écarts vont dans un sens différent d'un tableau à l'autre on risque donc de conclure à l'indépendance alors qu'il y a une dépendance au niveau de chaque tableau (erreur de seconde espèce). Cette situation correspond au cas où l'on a une interaction de niveau trois entre les facteurs. Ce test est donc à utiliser avec précaution.

Le test de Cochran-Mantel-Haenszel a été généralisé par Birch (1965), Landis et al. (1978) et Mantel et Byar (1978) au cas de tableaux LxC où L et C peuvent être plus grands que 2. Le calcul de M^2 est plus complexe mais aboutit toujours à une statistique qui suit asymptotiquement une loi du χ^2 à $(L-1)(C-1)$ degrés de liberté.

Il est recommandé de réaliser en parallèle du test CMH une analyse des V de Cramer pour les différents tableaux de contingence afin d'avoir une idée de leur contribution respective à l'indépendance. XLSTAT affiche automatiquement pour chacun des tableaux de contingence, un tableau avec les V de Cramer, les χ^2 et les p-values correspondantes (exactes pour les tableaux 2x2 et asymptotiques pour les tableaux de dimension supérieures) lorsque cela est possible, c'est-à-dire lorsqu'aucune somme marginale n'est nulle.

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.

 : cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.

Onglet **Général** :

Tableaux de contingence : si le format de données sélectionné est « tableaux de contingence », sélectionnez les k tableaux de contingence, puis précisez la valeur de k en entrant le **nombre de strates**.

Variable 1 : si le format de données sélectionné est « variables », sélectionnez les données correspondant à la première variable qualitative utilisée pour construire les tableaux de contingence.

Variable 2 : si le format de données sélectionné est « variables », sélectionnez les données correspondant à la seconde variable qualitative utilisée pour construire les tableaux de contingence.

Strates : si le format de données sélectionné est « variables », sélectionnez les données correspondant aux différentes strates.

Format des données : choisissez le format des données.

- **Tableaux de contingence** : activez cette option si vos données sont disponibles sous forme de k tableaux de contingence les uns en dessous des autres.
- **Variable s** : activez cette option si vos données se présentent sous la forme de deux variables qualitatives avec une ligne par individu, et d'une variable identifiant la strate pour chaque individu.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des colonnes/lignes : activez cette option si la première ligne (mode colonnes) ou colonne (mode lignes) des données sélectionnées contient des libellés.

Onglet **Options** :

Niveau de signification (%) : entrez la valeur du niveau de signification pour le test (valeur par défaut : 5%).

p-values exactes : activez cette option si vous souhaitez que XLSTAT calcule la p-value exacte dans la mesure du possible (voir description).

Hypothèse alternative : choisissez l'hypothèse alternative à utiliser pour le test (voir description) dans le cas d'un calcul de la p-value exacte sur tableaux 2x2.

Correction de continuité : activez cette option si vous souhaitez que XLSTAT utilise la correction de continuité si le calcul d'une p-value exacte n'est pas demandé ou s'il n'est pas possible (voir description).

Odds ratio commun : entrez la valeur de l'Odds ratio commun supposé pour les tableaux de contingence.

Onglet **Données manquantes** :

Ne pas accepter les valeurs manquantes : activez cette option pour que XLSTAT empêche la poursuite des calculs si des valeurs manquantes sont détectées.

Résultats

XLSTAT affiche les résultats du test, ainsi correspondent aux différentes statistiques des tests sélectionnés, et à l'interprétation qui en découle.

Exemple

Un exemple de test de Cochran-Mantel-Haenszel est disponible sur le Centre d'aide XLSTAT :

http://www.xlstat.com/demo-cmhf.htm

Bibliographie

Agresti A. (2002). Categorical Data Analysis, 2-nd Edition. John Wiley and Sons, New York.

Birch M. W. (1965). The detection of partial association II: the general case. Journal Roy Stat Soc B, **27**, 111-124.

Cochran W.G. (1954). Some methods for strengthening the common chi- squared tests. *Biometrics*, **10**, 417-451.

Hollander M. and Wolfe D. A. (1999). Nonparametric Statistical Methods, Second Edition. John Wiley and Sons, New York.

Landis J.R.., Heman E.R., Koch G.G. (1978). Average partial association in three way contingency tables: a review and discussion of alternative tests. Int Stat Rev., **46**, 237-354 (1978).

Mantel N. and Haenszel W. (1959) Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, **22**, 719-748.

Mantel N. and Byar D.P. (1978). Marginal homogeneity, symmetry and independence. *Communications in Statistics - Theory and Methods*, **A7**, 953-976 (1978).

Mehta C. R., Patel N. R., and Gray R. (1985). Computing an exact confidence interval for the common odds ratio in several 2 x 2 contingency tables. *Journal of the American Statistical Association*, **80**, 969-973.

Test des séquences pour un échantillon

Utilisez cet outil pour tester si une série d'événements binaires peut être considérée comme distribuée aléatoirement ou non.

Dans cette section :

Description

La première version de ce test non paramétrique a été présentée par Mood (1940) et utilise la même statistique que celle du test de comparaison de deux échantillons de Wald and Wolfowitz (1940), ce qui peut expliquer que ce test soit parfois mentionné par erreur sous le nom de test de Wald-Wolfowitz. Pour ajouter à la confusion, l'article de Mood fait plusieurs fois référence à l'article de Wald and Wolfowitz, notamment pour le calcul de la distribution asymptotique suivi par la statistique.

On définit par séquence une série d'événements identiques, précédés ou suivis par aucun événement ou des événements différents. Le test proposé par XLSTAT ne s'applique qu'à des événements binaires. Par exemple, pour ABBABBB, nous avons 4 séquences (A, BB, A, BBB).

XLSTAT accepte comme données d'entrée des données continues (binaires ou non) et des données catégorielles binaires. Pour les données continues, un point de séparation doit être choisi, afin que les données puissant être transformées en données binaires.

Un échantillon sera considéré comme aléatoirement distribué si aucune structure particulière ne peut être identifiée. Les cas extrêmes sont la répulsion (les deux événements sont à l'opposée dans la série), et l'intrication (les événements sont aussi alternés que possible). Avec l'exemple cité précédemment pour le cas de répulsion il y a "AABBBBB" ou "BBBBBAA", et pour l'intrication "BABABBB" ou "BABBABB" ou "BBABABB" ou "BBABBAB" ou encore "BBBABAB".

Dans le cas d'un test bilatéral, les hypothèses nulle (H0) et alternative (Ha) sont les suivantes :

- H0 : Les données sont distribuées au hasard.
- Ha : Les données ne sont pas distribuées au hasard.

Dans le cas unilatéral, il faut distinguer le test unilatéral à gauche et le test unilatéral à droite. Dans le test unilatéral à gauche, les hypothèses sont les suivantes :

- H0 : Les données sont distribuées au hasard.
- Ha : Il y a répulsion entre les deux types d'événements

Dans le test unilatéral à droite, les hypothèses sont les suivantes :

- H0 : Les données sont distribuées au hasard.
- Ha : Il y a intrication entre les deux types d'événements.

L'espérance du nombre de séquences R est:

$$E(R) = \frac{2mn}{N}$$

où m correspond au nombre d'événements du premier type, n au nombre d'événements du second type, and N est la somme de m et n.

La variance du nombre de séquences R est définie par :

$$V(R) = \frac{2mn(2mn - N)}{N^2(N-1)}$$

La valeur minimale possible de R est toujours 2. La valeur maximale est donnée par $2Min(m,n) - t$, où t est 1 si $n = m$, et 0 sinon.

Si r est le nombre de séquences observé sur l'échantillon, il a été montré par Wald et Wolfowitz qu'asymptotiquement, lorsque m ou n tendent vers l'infinie, on a

$$\frac{(r - E(R))}{\sqrt{V(R)}} \rightarrow N(0,1)$$

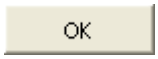
où $N(0,1)$ est la loi normale centrée réduite.

XLSTAT offre trois possibilités pour le calcul des p-value. Vous pouvez calculer la p-value à partir :

- de la distribution exacte de R,
- de la distribution asymptotique de R,
- d'une distribution approchée, calculée à partir de P permutations Monte Carlo. Comme le nombre de permutations possibles est égal à $N!$, P doit être fixé à une valeur élevée pour que l'approximation soit correcte.

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas (mode colonnes), XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite (mode lignes), XLSTAT considère que les variables sont en lignes et les observations en colonnes.

Onglet **Général**:

Données : sélectionnez la colonne (ou la ligne en mode ligne) de la série de données à analyser

Type de données :

- **Quantitative :** activez cette option pour sélectionner une colonne (ou une ligne en mode ligne) de données quantitatives. Les données seront alors transformées en fonction du point de séparation (voir plus bas).
- **Qualitative :** activez cette option pour sélectionner une colonne (ou une ligne en mode ligne) de données binaires.

Séparation : choisissez la valeur du point de séparation utilisé pour discrétiser les données continues en deux modalités.

- **Moyenne :** les observations sont séparées en fonction de la comparaison de leur valeur à la moyenne de l'échantillon.
- **Médiane :** les observations sont séparées en fonction de la comparaison de leur valeur à la médiane de l'échantillon.
- **Défini par utilisateur :** choisissez cette option pour transformer les données en fonction d'une valeur à saisir.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des colonnes/lignes : activez cette option si des en-têtes de colonne/ligne ont été sélectionnés.

Onglet **Options**:

Hypothèse alternative : choisissez l'hypothèse alternative à utiliser pour le test (voir description).

Niveau de signification (%) : entrez la valeur du niveau de signification pour les tests (valeur par défaut : 5%).

p-value exacte : activez cette option si vous souhaitez que XLSTAT calcule la p-value exacte (voir description).

p-value asymptotique : activez cette option si vous souhaitez que XLSTAT calcule la p-value exacte (voir description).

Correction de continuité : activez cette option si vous souhaitez que XLSTAT utilise la correction de continuité.

Méthode Monte Carlo : activez cette option si vous souhaitez que XLSTAT calcule la p-value approchée à partir de simulations Monte Carlo (voir description). Entrez alors le nombre de permutations aléatoires à réaliser.

Onglet **Données manquantes** :

Ne pas accepter les valeurs manquantes : activez cette option pour que XLSTAT empêche la poursuite des calculs si des valeurs manquantes sont détectées.

Supprimer les observations : activez cette option pour supprimer les observations comportant des données manquantes.

Résultats

Les résultats qui correspondent au test des séquences sont ensuite affichés. Une interprétation du test est fournie.

Bibliographie

Mood A. M. (1940). The distribution theory of runs. *Ann. Math. Statist.* , **11(4)**, 367-392.

Siegel S. and Castellan N. J. (1988). Nonparametric Statistics for the Behavioral Sciences, Second Edition. McGraw-Hill, New York, 58-54.

Wald A. and Wolfowitz J. (1940). On a test whether two samples are from the same population, *Ann. Math. Stat.*, **11(2)**, 147-162.

Test de Friedman-Rafsky

Utilisez cet outil pour comparer les distributions de deux échantillons décrits par des données quantitatives.

Dans cette section :

Description

Le test de Friedman-Rafsky est un test non-paramétrique sur deux échantillons X et Y. Ce test a pour hypothèse nulle :

$H0$: X et Y ont la même fonction de distribution i.e. $Fx = Fy$.

Il s'agit d'une généralisation du test de Wald-Wolfowitz en un test multivarié. La statistique de test dans le cas $p = 1$ (p étant le nombre de variables) se calcule comme suit :

- réunir les deux échantillons X de taille m et Y de taille n;
- trier les deux échantillons dans l'ordre croissant ;
- remplacer chaque nombre par X ou Y selon sa provenance ;
- compter r le nombre de X et de Y successifs ;
- enfin, calculer :

$$ \frac { (r - E (R)) } { \sqrt { V (R) } } \rightarrow N (0,1), $$

avec $E[R] = \frac{2mn}{N} + 1$, $Var(R) = \frac{2mn(2mn-m-n)}{N^2(N-1)}$ et $N = m + n$.

On rejette $H0$ pour des petites valeurs de r.

Dans le cas multivarié, il est moins simple de "trier" ses données. Pour y remédier, Friedman et Rafsky proposent d'utiliser un arbre couvrant minimum (en anglais *Minimum Spanning Tree*) qui remplace donc le tri des données. L'arbre couvrant minimum est un graphe non cyclique qui relie tous les sommets tout en ayant la plus petite somme de poids.

Le test procède comme suit :

- réunir les deux échantillons X de taille m et Y de taille n;
- calculer la matrice de distance ;
- construire l'arbre minimum couvrant en utilisant la matrice de distance comme un graphe complet ;

- compter r le nombre d'arêtes qui relient deux nœuds qui ne proviennent pas du même échantillon auquel on ajoute 1.

Le nombre r va permettre de calculer la statistique de test qui suit : $$ W = \frac{r - E[R]} {Var(R)^{\frac{1}{2}}} \rightarrow N (0,1), $$

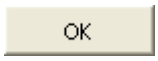
où $N(0,1)$ est la loi normale centrée réduite et avec : $$ E[R] = \frac{2mn}{N}+1, $$

et $$ Var(R) = \frac{2mn}{N(N-1)} { \frac{2mn-N}{N} + \frac{C-N+2}{(N-2)(N-3)}[N(N-1)-4mn+2]}, $$

où C est le nombre de paires d'arêtes qui partagent un nœud commun.

De même, on rejette $H0$ pour des petites valeurs de r.

Boîte de dialogue



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.


: cliquez sur ce bouton pour changer la façon dont vous chargez les données dans XLSTAT. Si le bouton possède une icône de souris, XLSTAT vous permet de faire une sélection à la souris des données. Si le bouton possède une icône de liste, XLSTAT vous permet de sélectionner les données à partir d'une liste. Si le bouton possède une icône orange, des boutons complémentaires avec un point d'interrogation  sont affichés pour vous permettre d'importer les données à partir de fichiers.

Onglet **Général** :

Format des données :

- **Échantillons séparés** : activez cette option pour sélectionner deux tableaux (un échantillon de taille m et un deuxième de taille n) qui peuvent avoir un nombre de lignes différent mais en veillant à ce qu'ils aient le même nombre p de colonnes.

- **Échantillons regroupés** : sélectionnez un tableau comprenant $m+n$ observations avec p variables quantitatives. Un identifiant (binaire) d'échantillon permettant d'affecter chaque observation à un échantillon doit par ailleurs être sélectionné.

Distance : cette option vous permet de sélectionner le calcul de distance que vous souhaitez utiliser suivant le type des variables :

- **distance euclidienne** ;
- **distance de Manhattan** ;
- **distance de Chebychev** ;
- **distance de Canberra**.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des variables : activez cette option si la première ligne des données sélectionnées (variables, libellés des observations, poids des observations) contient un libellé.

Arbre couvrant minimum / Algorithme : cette option vous permet de choisir entre trois méthodes pour construire l'arbre couvrant minimum :

- **Chazelle (Soft-Heap)** (par défaut) : cet algorithme est un algorithme déterministe et possède jusqu'à présent la plus petite complexité pour construire un arbre couvrant minimum. Cette complexité est due grâce à l'utilisation d'une "Soft-Heap" ("heap" se traduit par un "tas" ou une "pile"). Le temps d'exécution est estimé à $O(m\alpha(m,n))$ avec α la fonction inverse classique de la fonction d'Ackermann.
- **Kruskal** : l'algorithme de Kruskal est un des algorithmes le plus utilisé pour construire un arbre couvrant minimum. C'est un algorithme glouton qui est à privilégier pour des petits échantillons.
- **Boruvka** : il s'agit du premier algorithme inventé pour la construction d'un arbre couvrant minimum. C'est aussi un algorithme glouton.

Niveau de signification (%) : entrez la valeur du niveau de signification pour les tests (valeur par défaut : 5%).

Onglet **Données manquantes** :

Ne pas accepter les valeurs manquantes : activez cette option pour que XLSTAT empêche la poursuite des calculs si des valeurs manquantes sont détectées.

Supprimer les observations : activez cette option pour supprimer les observations comportant des données manquantes.

Estimer les données manquantes : activez cette option pour estimer les données manquantes avant le début des calculs.

- **Moyenne** : activez cette option pour estimer les données manquantes en utilisant la moyenne pour les variables correspondantes.

Onglet **Sorties** :

Statistiques descriptives : activez cette option pour afficher les statistiques descriptives pour les variables sélectionnées.

Arbre minimum couvrant : activez cette option pour afficher sur chaque ligne une branche de l'arbre, les deux noeuds qu'il relie, le poids et si les noeuds proviennent du même échantillon.

Matrice de distance : activez cette option pour afficher la matrice de distance.

Résultats

Statistiques descriptives : le tableau des statistiques descriptives présente des statistiques simples pour les données sélectionnées. Le nombre d'observations par variable et par échantillon, le minimum, le maximum, les différents quartiles, la moyenne, la variance et l'écart-type sont affichés.

Résultats associés au test de Friedman-Rafsky : le tableau montre les résultats détaillés du test, comme la valeur de la statistique W.

Résultats associés à l'arbre minimum couvrant : ce tableau comprend quatre colonnes qui informent sur : les arêtes de l'arbre, les deux noeuds qu'il relie, le poids (la distance entre les deux noeuds) et si les noeuds proviennent du même échantillon.

Résultat associé à la matrice de distance : dans ce tableau sont affichées les dissimilarités entre les objets des deux échantillons pour la distance choisie.

Exemple

Un exemple d'utilisation du test de Friedman-Rafsky est disponible sur le Centre d'aide XLSTAT :

https://www.xlstat.com/demo/rfy_en

Bibliographie

Friedman, J. H., & Rafsky, L. C. (1979). Multivariate generalizations of the Wald-Wolfowitz and Smirnov two-sample tests. *The Annals of Statistics*, 697-717.

Chazelle, B. (1997). A faster deterministic algorithm for minimum spanning trees. In Proceedings 38th Annual Symposium on Foundations of Computer Science (pp. 22-31). IEEE.

Chazelle, B. (2000). A minimum spanning tree algorithm with inverse-Ackermann type complexity. *Journal of the ACM* (JACM), 47(6), 1028-1047.

Chazelle, B. (2000). The soft heap: an approximate priority queue with optimal error rate. *Journal of the ACM* (JACM), 47(6), 1012-1027.

Tests pour les valeurs extrêmes

Test de Grubbs

Utilisez cet outil pour tester si une ou deux valeurs extrêmes (ou aberrantes) sont présentes dans un échantillon dont on suppose que la population dont il est extrait suit une loi normale.

Dans cette section :

Description

Les tests de Grubbs (1950, 1969, 1972) ont été mis au point pour permettre de déterminer si la valeur la plus grande, la valeur la plus petite, la valeur la plus grande ou la plus petite, ou dans le cas du test de Grubbs double, si les deux valeurs les plus grandes, ou si les deux plus petites peuvent être considérées comme extrêmes (ou aberrantes). Ce test suppose que les données correspondent à un échantillon provenant d'une population qui suit une loi normale.

Valeurs extrêmes

On appelle valeur extrême (ou aberrante) une donnée observée pour une variable qui semble anormale au regard des valeurs dont on dispose pour les autres observations de l'échantillon. On distingue deux types de situation dans lesquelles on rencontre des valeurs extrêmes :

- Une valeur extrême peut indiquer une erreur de lecture, une erreur de saisie ou un événement particulier qui a perturbé le phénomène observé au point de le rendre incomparable aux autres. Dans de tels cas, il faut soit corriger la valeur extrême si c'est possible, ou sinon supprimer l'observation.
- Une valeur extrême peut également être liée à un événement atypique, mais néanmoins connu ou intéressant à étudier. Par exemple, si l'on étudie la présence de certaines bactéries dans de l'eau de rivière, on peut avoir des prélèvements sans aucune bactérie, et d'autres avec des agrégats importants ou très importants. Ces données sont bien entendu importantes à conserver. Les modèles utilisés doivent alors tenir compte de cette dispersion possible.

Lorsque l'on rencontre des valeurs extrêmes, en fonction du stade de l'étude on doit, identifier les valeurs extrêmes, éventuellement à l'aide de tests, les marquer dans les rapports (tableaux ou graphiques), les supprimer ou utiliser des méthodes capables de les traiter comme tels.

Pour identifier les valeurs extrêmes, il existe différentes approches. Par exemple, en régression linéaire classique, on peut utiliser la valeur des D de Cook, ou soumettre les résidus standardisés au test de Grubbs ou de Dixon afin de voir si une ou deux valeurs sont anormales. Les tests de Grubbs et Dixon permettent d'identifier une ou deux valeurs aberrantes. Il est déconseillé d'utiliser itérativement ces méthodes sur un même échantillon, néanmoins cela peut être pertinent si l'on soupçonne réellement qu'il y a plus de deux valeurs extrêmes.

Définitions

Soit $x_1, x_2, \ldots, x_i, \ldots, x_n,$ un échantillon provenant d'une population que l'on suppose suivre une loi normale $N(\mu, s^2)$. Les paramètres μ et s^2 sont estimés par :

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

On définit :

$$x_{max} = \max_{i=1\ldots n}(x_i)$$

et

$$x_{min} = \min_{i=1\ldots n}(x_i)$$

Test de Grubbs (simple)

Les statistiques utilisées pour le test de Grubbs simple sont :

- Cas unilatéral à gauche : $G_{\min} = \frac{\overline{x} - x_{\min}}{s}$
- Cas unilatéral à droite : $G_{\max} = \frac{x_{\max} - \overline{x}}{s}$
- Cas bilatéral : $G = \max(G_{\min}, G_{\max})$

Dans le cas d'un test bilatéral, les hypothèses nulle (H0) et alternative (Ha) sont les suivantes :

- H0 : Il n'y a pas de valeur extrême dans l'échantillon.
- Ha : La valeur la plus petite ou la plus grande est une valeur extrême.

Dans le cas unilatéral, il faut distinguer le test unilatéral à gauche et le test unilatéral à droite. Dans le test unilatéral à gauche, les hypothèses sont les suivantes :

- H0 : Il n'y a pas de valeur extrême dans l'échantillon.
- Ha : La valeur la plus petite est une valeur extrême.

Dans le test unilatéral à droite, les hypothèses sont les suivantes :

- H0 : Il n'y a pas de valeur extrême dans l'échantillon.
- Ha : La valeur la plus grande est une valeur extrême.

On peut calculer une approximation G_{crit} de la valeur critique au-delà de laquelle, pour un niveau de signification α donné, on ne peut pas conserver l'hypothèse nulle. Elle est donnée par :

$$G_{crit}(n,\alpha) \approx \frac{(n-1)t_{n-2,1-\alpha/k}}{\sqrt{n-2+t^2_{n-2,1-\alpha/k}}}$$

où $t_{n-2,1-\alpha/k}$ est la valeur de la fonction de répartition inverse de Student pour $1-\alpha/k$ avec $n-2$ degrés de liberté et où k vaut n pour les tests unilatéraux et $2n$ pour le test bilatéral. On peut comparer cette valeur à celle de la statistique obtenue et en déduire que l'on peut conserver H0 si la G_{crit} est supérieure à G (ou G_{min} ou G_{max}) et la rejeter sinon. De l'approximation de G_{crit} on peut déduire par approximation la p-value associée à la valeur de la statistique obtenue. XLSTAT fournit l'ensemble de ces résultats ainsi que la conclusion du test en fonction du niveau de signification choisi par l'utilisateur.

Test de Grubbs double

On suppose ici que les observations x_i sont classées en ordre croissant. Les statistiques utilisées pour le test de Grubbs double sont :

- Cas unilatéral à gauche :

$$G2_{min} = \frac{Q_{min}}{(n-1)s}$$

avec

$$Q_{\min} = \sum_{i=3}^{n}(x_i - \overline{x}_3)^2, \overline{x}_3 = \frac{1}{n-2}\sum_{i=3}^{n} x_i$$

- Cas unilatéral à droite :

$$G2_{max} = \frac{Q_{max}}{(n-1)s}$$

avec

$$Q_{max} = \sum_{i=1}^{n-2}(x_i - \overline{x}_{n-2})^2, \overline{x}_{n-2} = \frac{1}{n-2}\sum_{i=1}^{n-2} x_i$$

- Cas bilatéral :

$$G2_{minmax} = \max(G2_{\min}, G2_{\max})$$

Dans le cas d'un test bilatéral, les hypothèses nulle (H0) et alternative (Ha) sont les suivantes :

- H0 : Il n'y a pas de valeur extrême dans l'échantillon.
- Ha : Les deux valeurs les plus petites ou les deux plus grandes sont des valeurs extrêmes.

Dans le test unilatéral à gauche, les hypothèses sont les suivantes :

- H0 : Il n'y a pas de valeur extrême dans l'échantillon.
- Ha : Les deux valeurs les plus petites sont des valeurs extrêmes.

Dans le test unilatéral à droite, les hypothèses sont les suivantes :

- H0 : Il n'y a pas de valeur extrême dans l'échantillon.
- Ha : Les deux valeurs les plus grandes sont des valeurs extrêmes.

La littérature (Wilrich, 2013) fournit une approximation $G2_{crit}$ de la valeur critique au-delà de laquelle, pour un niveau de signification α donné, on ne peut pas conserver l'hypothèse nulle. Néanmoins XLSTAT fournit une approximation des valeurs critiques sur la base de simulations Monte Carlo. Le nombre de ces approximations est par défaut fixé à 1000000, ce qui permet déjà d'obtenir des valeurs plus fiables que celles fournies dans les articles historique de Grubbs. XLSTAT fournit également sur la base de ces mêmes simulations une p-value, ainsi que la conclusion du test en fonction du niveau de signification choisi par l'utilisateur.

Z-scores

Parmi les résultats affichés par XLSTAT pour vous aider à identifier les valeurs extrêmes figurent les z-scores, qui correspondent aux x_i standardisés :

$$z_i = \frac{x_i - \overline{x}}{s}, (i = 1, \ldots, n)$$

Le problème de ces scores est qu'une fois l'intervalle admissible fixé (typiquement -1.96 et 1.96 pour un intervalle à 95%), toute valeur qui se trouve en dehors est considérée comme suspecte. Hors on sait que si l'on a 100 valeurs, il est statistiquement normal d'en avoir 5 en dehors de cet intervalle. Par ailleurs, par construction le z-score le plus élevé vaut au maximum :

$$\max_{i=1\ldots n} z_i \leq \frac{n-1}{\sqrt{n}}$$

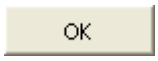
Iglewicz et Hoaglin (1993) recommandent l'utilisation d'un z-score modifié afin de mieux faire ressortir les valeurs extrêmes :

$$z_i = 0.6745\frac{x_i - \overline{x}}{MDA}, (i = 1, \ldots, n)$$

où MDA est la médiane des déviations absolues (*Median Absolute Deviation*). L'intervalle acceptable est]-3.5 ; 3.5[quelque soit n.

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas (mode colonnes), XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite (mode lignes), XLSTAT considère que les variables sont en lignes et les observations en colonnes.

Onglet **Général**:

Données : sélectionnez les données sur la feuille Excel. Si vous sélectionnez plusieurs colonnes, XLSTAT considère que chaque colonne correspond à un échantillon différent. Si des en-têtes ont été sélectionnés, veuillez vérifier que l'option « Libellés des colonnes » est activée.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des colonnes : activez cette option si la première ligne des données sélectionnées contient un libellé.

Vous pouvez choisir le type de test à appliquer sur les données :

- **Test de Grubbs** : choisissez ce test pour effectuer un test de Grubbs simple.
- **Test de Grubbs double** : choisissez ce test pour effectuer un test de Grubbs double.

Onglet **Options**:

Hypothèse alternative : choisissez l'hypothèse alternative à utiliser pour le test (voir description).

Niveau de signification (%) : entrez le niveau de signification à utiliser pour les différents tests (valeur par défaut : 5%).

Itérations : choisissez si vous ne voulez appliquer le test choisi sur vos données qu'un nombre limité de fois (par défaut 1 fois), ou si vous voulez laisser XLSTAT itérer jusqu'à ce que plus aucune donnée extrême ne soit trouvée.

Valeur critique / p-value : Entrez le nombre de simulations aléatoires à réaliser pour le calcul de la valeur critique et de la p-value, ainsi que le temps maximum en secondes. Cette option n'est disponible que pour le test de Grubbs double.

Onglet **Données manquantes** :

Ne pas accepter les valeurs manquantes : activez cette option pour que XLSTAT empêche la poursuite des calculs si des valeurs manquantes sont détectées.

Supprimer les observations : activez cette option pour supprimer les observations comportant des données manquantes.

Ignorer les données manquantes : activez cette option pour ignorer les données manquantes.

Onglet **Sorties**:

Statistiques descriptives : activez cette option pour calculer et afficher les statistiques descriptives pour les différents échantillons.

Z-scores : activez cette option pour calculer et afficher les z-scores et le graphique correspondant. Vous avez le choix entre les **z-scores modifiés** ou les **z-scores classiques**. Dans le cas de ces derniers vous pouvez choisir deux limites à afficher sur les graphiques.

Résultats

Statistiques descriptives : dans ce tableau sont affichées les statistiques descriptives correspondant aux différents traitements.

Les résultats qui correspondent au **test de Grubbs** sont ensuite affichés. Une interprétation du test est fournie si une seule itération du test a été demandée, ou si aucune observation n'a été identifiée comme extrême dès la première itération.

Dans le cas où plusieurs itérations ont été demandées, est également affiché un tableau donnant, pour chaque observation, l'itération au cours de laquelle elle a été retirée de l'échantillon.

Les z-scores sont ensuite affichés s'ils ont été demandés.

Exemple

Un exemple de test de Grubbs est disponible sur le Centre d'aide XLSTAT :

http://www.xlstat.com/demo-grubbsf.htm

Bibliographie

Barnett V. and Lewis T. (1980). Outliers in Statistical Data. John Wiley and Sons, Chichester, New York, Brisbane, Toronto.

Grubbs F.E. (1950). Sample criteria for testing outlying observations. *Ann. Math. Stat.* **21**, 27-58.

Grubbs F.E. (1969). Procedures for detecting outlying observations in samples. *Technometrics*, **11(1)**, 1-21.

Grubbs, F.E. and Beck G. (1972). Extension of sample sizes and percentage points for significance tests of outlying observations. *Technometrics*, **14**, 847-854.

Hawkins D.M. (1980). Identification of Outliers. Chapman and Hall, London.

Iglewicz B. and Hoaglin D. (1993). "Volume 16: How to Detect and Handle Outliers", The ASQC Basic References in Quality Control: Statistical Techniques, Edward F. Mykytka, Ph.D., Editor.

International Organization for Standardization (1994). ISO 5725-2: Accuracy (trueness and precision) of measurement methods and results—Part 2: Basic method for the determination of repeatability and reproducibility of a standard measurement method, Geneva.

Snedecor G. W. and Cochran W. G. (1989). Statistical Methods, Eighth Edition, Iowa State University Press.

Wilrich P. -T. (2013). Critical values of Mandel's h and k, the Grubbs and the Cochran test statistic. *Advances in Statistical Analysis*, 97(1), 1-10.

Test de Dixon

Utilisez cet outil pour tester si une ou deux valeurs extrêmes (ou aberrantes) sont présentes dans un échantillon dont on suppose que la population dont il est extrait suit une loi normale.

Dans cette section :

Description

Le test de Dixon (1950, 1951, 1953), qui est en réalité subdivisé en 6 tests en fonction de la statistique choisie et du nombre de valeurs extrêmes à identifier, a été mis au point pour permettre de déterminer si la valeur la plus grande ou la valeur la plus petite d'un échantillon, ou les deux valeurs les plus grandes, ou les deux plus petites peuvent être considérées comme extrêmes (ou aberrantes). Ce test suppose que les données correspondent à un échantillon provenant d'une population qui suit une loi normale.

Valeurs extrêmes

On appelle valeur extrême (ou aberrante) une donnée observée pour une variable qui semble anormale au regard des valeurs dont on dispose pour les autres observations de l'échantillon. On distingue deux types de situation dans lesquelles on rencontre des valeurs extrêmes :

- Une valeur extrême peut indiquer une erreur de lecture, une erreur de saisie ou un événement particulier qui a perturbé le phénomène observé au point de le rendre incomparable aux autres. Dans de tels cas, il faut soit corriger la valeur extrême si c'est possible, ou sinon supprimer l'observation.

- Une valeur extrême peut également être liée à un événement atypique, mais néanmoins connu ou intéressant à étudier. Par exemple, si l'on étudie la présence de certaines bactéries dans de l'eau de rivière, on peut avoir des prélèvements sans aucune bactérie, et d'autres avec des agrégats importants ou très importants. Ces données sont bien entendu importantes à conserver. Les modèles utilisés doivent alors tenir compte de cette dispersion possible.

Lorsque l'on rencontre des valeurs extrêmes, en fonction du stade de l'étude, on doit, identifier les valeurs extrêmes, éventuellement à l'aide de tests, les marquer dans les rapports (tableaux ou graphiques), les supprimer ou utiliser des méthodes capables de les traiter comme tels.

Pour identifier les valeurs extrêmes, il existe différentes approches. Par exemple, en régression linéaire classique, on peut utiliser la valeur des D de Cook, ou soumettre les résidus standardisés au test de Grubbs ou de Dixon afin de voir si une ou deux valeurs sont anormales. Les tests de Grubbs et Dixon permettent d'identifier une ou deux valeurs aberrantes. Il est déconseillé d'utiliser itérativement ces méthodes sur un même échantillon, néanmoins cela peut être pertinent si l'on soupçonne réellement qu'il y a plus de deux valeurs extrêmes.

Définitions

Soit un échantillon $x_1, x_2, \ldots, x_i, \ldots, x_n$, un échantillon de taille n provenant d'une population que l'on suppose suivre une loi normale $N(\mu, \sigma^2)$. Les paramètres μ et σ^2 sont estimés par :

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

et

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{(n-1)}$$

On suppose ici que les observations x_i sont classées en ordre croissant.

Test de Dixon pour une valeur extrême

Ce test est utilisé pour déterminer, si la valeur la plus grande ou la plus petite peut être considérée comme extrême (ou aberrante). Ce test suppose que les données correspondent à un échantillon provenant d'une population qui suit une loi normale.

Les statistiques utilisées pour le test de Dixon et les plages d'utilisation correspondantes validées dans la littérature (Barnett et Lewis 1994 et Verma et Quiroz-Ruiz 2006) sont :

- $R_{10} = \frac{x_n - x_{n-1}}{x_n - x_1}$, recommandée pour $3 \leq n \leq 100$, aussi nommée N7

- $R_{11} = \frac{x_n - x_{n-1}}{x_n - x_2}$, recommandée pour $4 \leq n \leq 100$, aussi nommée N9

- $R_{12} = \frac{x_n - x_{n-1}}{x_n - x_3}$, recommandée pour $5 \leq n \leq 100$, aussi nommée N10

Ces statistiques sont valables pour tester si la valeur maximale est une valeur extrême. Pour identifier si la valeur minimale est une valeur extrême, il suffit de trier les données en ordre décroissant et d'utiliser les mêmes statistiques. Si l'on veut identifier si le minimum ou le maximum est une valeur extrême, on fait le calcul de la statistique pour les deux alternatives (tri ascendant ou descendant) et on retient la valeur la plus grande.

Dans le cas d'un test bilatéral, les hypothèses nulle (H0) et alternative (Ha) sont les suivantes :

- H0 : Il n'y a pas de valeur extrême dans l'échantillon.
- Ha : La valeur la plus petite ou la plus grande est une valeur extrême.

Dans le cas unilatéral, il faut distinguer le test unilatéral à gauche et le test unilatéral à droite. Dans le test unilatéral à gauche, les hypothèses sont les suivantes :

- H0 : Il n'y a pas de valeur extrême dans l'échantillon.
- Ha : La valeur la plus petite est une valeur extrême.

Dans le test unilatéral à droite, les hypothèses sont les suivantes :

- H0 : Il n'y a pas de valeur extrême dans l'échantillon.
- Ha : La valeur la plus grande est une valeur extrême.

Test de Dixon pour deux valeurs extrêmes

Ce test est utilisé pour déterminer, si les deux valeurs les plus grandes ou les plus petites peuvent être considérées comme extrêmes (ou aberrantes). Ce test suppose que les données correspondent à un échantillon provenant d'une population qui suit une loi normale.

Les statistiques utilisées pour le test de Dixon pour deux valeurs extrêmes et les plages d'utilisation correspondantes validées dans la littérature (Barnett et Lewis 1994 et Verma et Quiroz-Ruiz 2006) sont :

- $R_{20} = \frac{x_n - x_{n-2}}{x_n - x_1}$, recommandée pour $4 \leq n \leq 100$, aussi nommée N11

- $R_{21} = \frac{x_n - x_{n-2}}{x_n - x_2}$, recommandée pour $5 \leq n \leq 100$, aussi nommée N12

- $R_{22} = \frac{x_n - x_{n-2}}{x_n - x_3}$, recommandée pour $6 \leq n \leq 100$, aussi nommée N13

Ces statistiques sont valables pour tester si la valeur maximale est une valeur extrême. Pour identifier si la valeur minimale est une valeur extrême, il suffit de trier les données en ordre décroissant et d'utiliser les mêmes statistiques. Si l'on veut identifier si le minimum ou le maximum est une valeur extrême, on fait le calcul de la statistique pour les deux alternatives (tri ascendant ou descendant) et on retient la valeur la plus grande.

Dans le cas d'un test bilatéral, les hypothèses nulle (H0) et alternative (Ha) sont les suivantes :

- H0 : Il n'y a pas de valeur extrême dans l'échantillon.
- Ha : Les deux valeurs les plus petites ou les deux plus grandes sont des valeurs extrêmes.

Dans le cas unilatéral, il faut distinguer le test unilatéral à gauche et le test unilatéral à droite. Dans le test unilatéral à gauche, les hypothèses sont les suivantes :

- H0 : Il n'y a pas de valeur extrême dans l'échantillon.
- Ha : Les deux valeurs les plus petites sont des valeurs extrêmes.

Dans le test unilatéral à droite, les hypothèses sont les suivantes :

- H0 : Il n'y a pas de valeur extrême dans l'échantillon.
- Ha : Les deux valeurs les plus grandes sont des valeurs extrêmes.

Calcul de la p-value et de la valeur critique pour le test de Dixon

La littérature fournit des approximations plus ou moins précises de la valeur critique au-delà de laquelle, pour un niveau de signification a donné, on ne peut pas conserver l'hypothèse nulle. Néanmoins XLSTAT fournit une approximation des valeurs critiques sur la base de simulations Monte Carlo. Le nombre de ces approximations est par défaut fixé à 1000000, ce qui permet d'obtenir des valeurs plus fiables que celles fournies dans les articles historiques de Dixon. XLSTAT fournit également sur la base de ces mêmes simulations une p-value, ainsi que la conclusion du test en fonction du niveau de signification choisi par l'utilisateur.

Z-scores

Parmi les résultats affichés par XLSTAT pour vous aider à identifier les valeurs extrêmes figurent les z-scores, qui correspondent aux xi standardisés :

$$z_i = \frac{x_i - \overline{x}}{s} \quad (i = 1, \ldots, n)$$

Le problème de ces scores est qu'une fois l'intervalle admissible fixé (typiquement -1.96 et 1.96 pour un intervalle à 95%), toute valeur qui se trouve en dehors est considérée comme suspecte. Hors on sait que si l'on a 100 valeurs, il est statistiquement normal d'en avoir 5 en dehors de cette intervalle. Par ailleurs, par construction le z-score le plus élevé vaut au maximum :

$$\max_{i=1\ldots n} z_i \leq \frac{n-1}{\sqrt{n}}$$

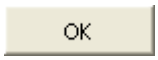
Iglewicz et Hoaglin (1993) recommandent l'utilisation d'un z-score modifié afin de mieux faire ressortir les valeurs extrêmes :

$z_i = 0.6745 \frac{x_i - \overline{x}}{MDA}$ (i=1,...n)

où MDA est la médiane des déviations absolues (*Median Absolute Deviation*). L'intervalle acceptable est]-3.5 ; 3.5[quelque soit n.

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas (mode colonnes), XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite (mode lignes), XLSTAT considère que les variables sont en lignes et les observations en colonnes.

Onglet **Général**:

Données : sélectionnez les données sur la feuille Excel. Si vous sélectionnez plusieurs colonnes, XLSTAT considère que chaque colonne correspond à un échantillon différent. Si des en-têtes ont été sélectionnés, veuillez vérifier que l'option « Libellés des colonnes » est activée.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des colonnes : activez cette option si la première ligne des données sélectionnées contient un libellé.

Vous pouvez choisir le type de test à appliquer sur les données :

- **Défini par l'utilisateur** : choisissez la statistique à utiliser pour effectuer un test de Dixon.
- **Automatique** : choisissez cette option pour laisser XLSTAT choisir la statistique à utiliser, suivant les recommandations de la littérature (Böhrer, 2008).

Onglet **Options**:

Hypothèse alternative : choisissez l'hypothèse alternative à utiliser pour le test (voir description).

Niveau de signification (%) : entrez le niveau de signification à utiliser pour les différents tests (valeur par défaut : 5%).

Itérations : choisissez si vous ne voulez appliquer le test choisi sur vos données qu'un nombre limité de fois (par défaut 1 fois), ou si vous voulez laisser XLSTAT itérer jusqu'à ce que plus aucune donnée extrême ne soit trouvée.

Valeur critique / p-value : Entrez le nombre de simulations aléatoires à réaliser pour le calcul de la valeur critique et de la p-value ainsi que le temps maximum en secondes.

Onglet **Données manquantes** :

Ne pas accepter les valeurs manquantes : activez cette option pour que XLSTAT empêche la poursuite des calculs si des valeurs manquantes sont détectées.

Supprimer les observations : activez cette option pour supprimer les observations comportant des données manquantes.

Ignorer les données manquantes : activez cette option pour ignorer les données manquantes.

Onglet **Sorties**:

Statistiques descriptives : activez cette option pour calculer et afficher les statistiques descriptives pour les différents échantillons.

Scores Z : activez cette option pour calculer et afficher les scores z et le graphique correspondant. Vous avez le choix entre les **scores z modifiés** ou les **scores z classiques**. Dans le cas de ces derniers vous pouvez choisir deux limites à afficher sur les graphiques.

Résultats

Statistiques descriptives : dans ce tableau sont affichées les statistiques descriptives correspondant aux différents échantillons.

Les résultats qui correspondent au **test de Dixon** sont ensuite affichés. Une interprétation du test est fournie si une seule itération du test a été demandée, ou si aucune observation n'a été identifiée comme extrême dès la première itération.

Dans le cas où plusieurs itérations ont été demandées, est également affiché un tableau donnant, pour chaque observation, l'itération au cours de laquelle elle a été retirée de l'échantillon.

Les z-scores sont ensuite affichés s'ils ont été demandés.

Exemple

Un exemple de test de Dixon est disponible sur le Centre d'aide XLSTAT :

http://www.xlstat.com/demo-dixonf.htm

Bibliographie

Bö hrer A. (2008). One-sided and Two-sided Critical Values for Dixon's Outlier Test for Sample Sizes up to n = 30. *Economic Quality Control* , **23(1)**, 5-13.

Barnett V. and Lewis T. (1980). Outliers in Statistical Data. John Wiley and Sons, Chichester, New York, Brisbane, Toronto.

Dixon W.J. (1950). Analysis of extreme values. *Annals of Math. Stat.*, **21**, 488-506.

Dixon W.J. (1951). Ratios involving of extreme values. *Annals of Math. Stat.*, **22**, 68-78.

Dixon W.J. (1953). Processing data for outliers. *J. Biometrics*, **9**, 74-89.

Hawkins D.M. (1980). Identification of Outliers. Chapman and Hall, London.

International Organization for Standardization (1994). ISO 5725-2: Accuracy (trueness and precision) of measurement methods and results—Part 2: Basic method for the determination of repeatability and reproducibility of a standard measurement method, Geneva.

Verma S. P. and Quiroz-Ruiz A. (2006). Critical values for six Dixon tests for outliers in normal samples up to sizes 100, and applications in science and engineering, Revista Mexicana de Ciencias Geológicas, **23(2)**, 133-161.

Test du C de Cochran

Utilisez cet outil pour tester si une variance est anormale parmi une série de k variances.

Dans cette section :

Description

Le test de Cochran (Cochran 1941) fait partie des tests développés pour permettre d'identifier et d'étudier l'homogénéité d'une série de variances (test de Bartlett, de Brown-Forsythe, de Levene ou de Hartley notamment). Le test de Cochran a été développé pour répondre à une question bien précise : les variances sont-elles homogènes ou la variance la plus élevée est-elle différente des autres ? XLSTAT propose également deux alternatives et utilise les résultats de 't Lam (2010) pour une extension du cas équilibré au cas déséquilibré, qui permet également une généralisation au test bilatéral.

Valeurs extrêmes

On appelle valeur extrême (ou aberrante) une donnée observée pour une variable qui semble anormale au regard des valeurs dont on dispose pour les autres observations de l'échantillon. On distingue deux types de situation dans lesquelles on rencontre des valeurs extrêmes :

- Une valeur extrême peut indiquer une erreur de lecture, une erreur de saisie ou un événement particulier qui a perturbé le phénomène observé au point de le rendre incomparable aux autres. Dans de tels cas, il faut soit corriger la valeur extrême si c'est possible, ou sinon supprimer l'observation.
- Une valeur extrême peut également être liée à un événement atypique, mais néanmoins connu ou intéressant à étudier. Par exemple, si l'on étudie la présence de certaines bactéries dans de l'eau de rivière, on peut avoir des prélèvements sans aucune bactérie, et d'autres avec des agrégats importants ou très importants. Ces données sont bien entendu importantes à conserver. Les modèles utilisés doivent alors tenir compte de cette dispersion possible.

Lorsque l'on rencontre des valeurs extrêmes, en fonction du stade de l'étude on doit, identifier les valeurs extrêmes, éventuellement à l'aide de tests, les marquer dans les rapports (tableaux ou graphiques), les supprimer ou utiliser des méthodes capables de les traiter comme tels.

Pour identifier les valeurs extrêmes, il existe différentes approches. Par exemple, en régression linéaire classique, on peut utiliser la valeur des D de Cook, ou soumettre les résidus standardisés au test de Grubbs ou de Dixon afin de voir si une ou deux valeurs sont anormales. Les tests de Grubbs et Dixon permettent d'identifier une ou deux valeurs aberrantes. Il est déconseillé d'utiliser itérativement ces méthodes sur un même échantillon, néanmoins cela peut être pertinent si l'on soupçonne réellement qu'il y a plus de deux valeurs extrêmes.

Si l'échantillon peut être subdivisé en sous-échantillons, on peut s'intéresser aux variations d'un sous-échantillon à l'autre. Le test du C de Cochran et les statistiques h et k de Mandel font partie des méthodes adaptées à ce type d'études.

Définitions

Soit un échantillon $x_{11}, x_{12}, \ldots, x_{1n1}, x_{12}, x_{22}, \ldots x_{2n2}, \ldots, x_{p1}, x_{p2}, \ldots, x_{pnp},$ de données que l'on distingue pour leur appartenance à p groupes (par exemple des laboratoires) d'effectif respectif $n_i (i = 1 \ldots p)$. Soit $\overline{x}_i$ la moyenne de l'échantillon correspondant au groupe i et s_i^2 la variance estimée pour le groupe i. On a :

$$\overline{x}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij}$$

$$s_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (x_{ij} - \overline{x_i})^2$$

On suppose ici que les observations sont identiquement distribuées suivant une loi normale.

Test du C de Cochran

La statistique C_i pour le groupe i $(i = 1 \ldots p)$ donnée par Cochran (1941) est :

$$C_i = \frac{s_i^2}{\sum_{j=1}^{p} s_j^2}$$

et

$$C = \max_{i=1\ldots p}(C_i)$$

est la statistique utilisée pour le test. La valeur critique correspondant à cette statistique a été abondamment tabulée et différents auteurs ont fourni des approximations (Wilrich, 2013). Cependant, comme le note 't Lam (2010), cette statistique présente des limites :

- pour être correct, ce test suppose que les groupes sont d'effectifs égaux (plan équilibré),
- seule la variance maximale est étudiée, la variance minimale étant négligée même si c'est à son niveau que se trouve l'anomalie (test unilatéral à droite uniquement),

- les tables de valeurs critiques sont limitées et comportent parfois des erreurs
- l'utilisation de tables n'est pas particulièrement pratique.

Pour cette raison, 't Lam propose une généralisation de la statistique de Cochran pour les plans déséquilibrés, et une généralisation à un test unilatéral à gauche ou bilatéral. La statistique pour le groupe i est donnée par :

$$G_i = \frac{\nu_i S_i^2}{\sum_{j=1}^{p} \nu_j s_j^2} \text{ avec } \nu_i = n_i - 1$$

Pour un niveau de signification a donné, 't Lam donne les valeurs critiques inférieures et supérieures pour cette statistique.

$$G_{LL}(i) = \left[1 + \frac{(\nu_{total}/\nu_i) - 1}{F^{-1}(\delta/p, \nu_i, \nu_{total} - \nu_i)}\right]^{-1} \tag{1}$$

et

$$G_{UL}(i) = \left[1 + \frac{(\nu_{total}/\nu_i) - 1}{F^{-1}(1 - \delta/p, \nu_i, \nu_{total} - \nu_i)}\right]^{-1} \tag{2}$$

avec $\nu_{total} = (\sum_{i=1}^{p} \nu_i) - p$, et $\delta = \alpha$ pour un test unilatéral et $\delta = \alpha/2$ pour un test bilatéral, et F^{-1} est la fonction de répartition inverse de la loi de Fisher.

Dans le cas d'un test bilatéral, les hypothèses nulle (H_0) et alternative (H_a) sont les suivantes :

- H_0 : Les variances de l'échantillon sont homogènes.
- H_a : Au moins l'une des variances est différente des autres.

Dans le cas unilatéral, il faut distinguer le test unilatéral à gauche et le test unilatéral à droite. Dans le test unilatéral à gauche, les hypothèses sont les suivantes :

- H_0 : Les variances de l'échantillon sont homogènes.
- H_a : Au moins l'une des variances est inférieure aux autres.

Dans le test unilatéral à droite, les hypothèses sont les suivantes :

- H_0 : Les variances de l'échantillon sont homogènes.
- H_a : Au moins l'une des variances est supérieure aux autres.

Dans le cadre d'un test bilatéral, pour identifier la variance potentiellement extrême on calcule :

$$Gmin = \min_{i=1...k}(G_i) \text{ et } Gmax = \max_{i=1...k}(G_i)$$

Puis, si l'une ou les deux statistiques ne sont pas dans l'intervalle critique donné par (1) et (2), on calcule les p-values associées aux deux statistiques. On identifiera comme extrême la variance associée à la statistique donnant la p-value la plus faible.

Z-scores

Parmi les résultats affichés par XLSTAT pour vous aider à identifier les valeurs extrêmes figurent les z-scores, qui correspondent aux x_i standardisés :

$$z_i = \frac{x_i - \overline{x}}{s} \ (i = 1, ...n)$$

Le problème de ces scores est qu'une fois l'intervalle admissible fixé (typiquement -1.96 et 1.96 pour un intervalle à 95%), toute valeur qui se trouve en dehors est considérée comme suspecte. Hors on sait que si l'on a 100 valeurs, il est statistiquement normal d'en avoir 5 en dehors de cette intervalle. Par ailleurs, par construction le z-score le plus élevé vaut au maximum :

$$\max_{i=1...n} z_i \leq \frac{n-1}{\sqrt{n}}$$

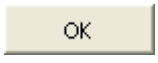
Iglewicz et Hoaglin (1993) recommandent l'utilisation d'un z-score modifié afin de mieux faire ressortir les valeurs extrêmes :

$$z_i = 0.6745\frac{x_i - \overline{x}}{MDA} \ (i = 1, ...n)$$

où MDA est la médiane des déviations absolues (*Median Absolute Deviation*). L'intervalle acceptable est]-3.5 ; 3.5[quelque soit n.

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.

 : cliquez sur ce bouton pour effacer les sélections de données.

 : cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas (mode colonnes), XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite (mode lignes), XLSTAT considère que les variables sont en lignes et les observations en colonnes.

Onglet **Général** :

Données : Si le format de données sélectionné est « une colonne par variable », sélectionnez les données pour les différentes variables. Si le format de données sélectionné est « une colonne par groupe », sélectionnez les échantillons associés aux différents groupes. Si le format de données sélectionné est « variances », sélectionnez les variances de chaque groupe.

Identifiant de groupe / Taille de groupe : si le format de données sélectionné est « une colonne par variable », sélectionnez les données identifiant les groupes auxquels les données sélectionnées correspondent. Si le format de données sélectionné est « Variances », vous devez alors entrer la **taille de groupe** si le plan est équilibré, ou sélectionner les tailles des groupes si le plan est déséquilibré.

Format des données : choisissez le format des données.

- **Une colonne/ligne par groupe** : activez cette option pour sélectionner une colonne (ou ligne en mode lignes) par groupe.
- **Une colonne/ligne par variable** : activez cette option pour que XLSTAT fasse autant de tests qu'il y a de colonnes (ou lignes en mode lignes), sachant que chaque colonne/ligne doit contenir le même nombre de lignes/colonnes, et qu'un identifiant de groupe doit par ailleurs être sélectionné.
- **Variances** : activez cette option si vos données correspondent à des variances déjà calculées. Vous devez alors entrer la taille de groupe si le plan est équilibré, ou sélectionner les tailles des groupes si le plan est déséquilibré.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des colonnes : activez cette option si la première ligne des données sélectionnées contient un libellé.

Vous pouvez choisir le type de test à appliquer sur les données :

- **C de Cochran (équilibré)** : choisissez cette option si le plan est équilibré et que vous souhaitez réaliser un test unilatéral.
- **G de 't Lam** : choisissez cette option si le plan est déséquilibré et/ou si vous souhaitez réaliser un test bilatéral.

Onglet **Options** :

Hypothèse alternative : choisissez l'hypothèse alternative à utiliser pour le test (voir description).

Niveau de signification (%) : entrez le niveau de signification à utiliser pour les différents tests (valeur par défaut : 5%).

Itérations : choisissez si vous ne voulez appliquer le test choisi sur vos données qu'un nombre limité de fois (par défaut 1 fois), ou si vous voulez laisser XLSTAT itérer jusqu'à ce que plus aucune donnée extrême ne soit trouvée.

Onglet **Données manquantes** :

Ne pas accepter les valeurs manquantes : activez cette option pour que XLSTAT empêche la poursuite des calculs si des valeurs manquantes sont détectées.

Supprimer les observations : activez cette option pour supprimer les observations comportant des données manquantes.

Onglet **Sorties** :

Statistiques descriptives : activez cette option pour calculer et afficher les statistiques descriptives pour les différents échantillons.

Scores z : activez cette option pour calculer et afficher les scores z et le graphique correspondant. Vous avez le choix entre les **scores z modifiés** ou les **scores z classiques**. Dans le cas de ces derniers vous pouvez choisir deux limites à afficher sur les graphiques.

Résultats

Statistiques descriptives : dans ce tableau sont affichées les statistiques descriptives correspondant aux différents traitements.

Les résultats qui correspondent au **test de Cochran** sont ensuite affichés. Une interprétation du test est fournie si une seule itération du test a été demandée, ou si aucune observation n'a été identifiée comme extrême dès la première itération.

Dans le cas où plusieurs itérations ont été demandées, est également affiché un tableau donnant, pour chaque observation, l'itération au cours de laquelle elle a été retirée de

l'échantillon.

Les scores z sont ensuite affichés s'ils ont été demandés.

Exemple

Un exemple de test de Cochran est disponible sur le Centre d'aide XLSTAT :

http://www.xlstat.com/demo-cochran-cf.htm

Bibliographie

Cochran W.G. (1941). The distribution of the largest of a set of estimated variances as a fraction of their total. *Ann. Eugen.* **11**, 47-52.

Barnett V. and Lewis T. (1980). Outliers in Statistical Data. John Wiley and Sons, Chichester, New York, Brisbane, Toronto.

Hawkins D.M. (1980). Identification of Outliers. Chapman and Hall, London.

Iglewicz B. and Hoaglin D. (1993). "Volume 16: How to Detect and Handle Outliers", The ASQC Basic References in Quality Control: Statistical Techniques, Edward F. Mykytka, Ph.D., Editor.

International Organization for Standardization (1994). ISO 5725-2: Accuracy (trueness and precision) of measurement methods and results—Part 2: Basic method for the determination of repeatability and reproducibility of a standard measurement method, Geneva.

't Lam R.U.E. (2010). Scrutiny of variance results for outliers: Cochran's test optimized? *Analytica Chimica Acta*, **659**, 68-84.

Wilrich P. -T. (2013). Critical values of Mandel's h and k, the Grubbs and the Cochran test statistic. *Advances in Statistical Analysis*, 97(1), 1-10.

Statistiques h et k de Mandel

Utilisez cet outil pour calculer les statistiques h et k de Mandel pour un échantillon afin d'identifier des valeurs extrêmes.

Dans cette section :

Description

Les statistiques h et k de Mandel (1985, 1991) on été développées afin d'identifier dans le cadre d'analyses inter-laboratoires des laboratoires non conformes.

Valeurs extrêmes

On appelle valeur extrême (ou aberrante) une donnée observée pour une variable qui semble anormale au regard des valeurs dont on dispose pour les autres observations de l'échantillon. On distingue deux types de situation dans lesquelles on rencontre des valeurs extrêmes :

- Une valeur extrême peut indiquer une erreur de lecture, une erreur de saisie ou un événement particulier qui a perturbé le phénomène observé au point de le rendre incomparable aux autres. Dans de tels cas, il faut soit corriger la valeur extrême si c'est possible, ou sinon supprimer l'observation.
- Une valeur extrême peut également être liée à un événement atypique, mais néanmoins connu ou intéressant à étudier. Par exemple, si l'on étudie la présence de certaines bactéries dans de l'eau de rivière, on peut avoir des prélèvements sans aucune bactérie, et d'autres avec des agrégats importants ou très importants. Ces données sont bien entendu importantes à conserver. Les modèles utilisés doivent alors tenir compte de cette dispersion possible.

Lorsque l'on rencontre des valeurs extrêmes, en fonction du stade de l'étude on doit, identifier les valeurs extrêmes, éventuellement à l'aide de tests, les marquer dans les rapports (tableaux ou graphiques), les supprimer ou utiliser des méthodes capables de les traiter comme tels.

Pour identifier les valeurs extrêmes, il existe différentes approches. Par exemple, en régression linéaire classique, on peut utiliser la valeur des D de Cook, ou soumettre les résidus standardisés au test de Grubbs afin de voir si une ou deux valeurs sont anormales. Le test de Grubbs simple permet d'identifier une valeur aberrante, le test de Grubbs double permet d'en

identifier deux. Il est déconseillé d'utiliser itérativement ces méthodes sur un même échantillon, néanmoins cela peut être pertinent si l'on soupçonne réellement qu'il y a plus de deux valeurs extrêmes.

Si l'échantillon peut être subdivisé en sous-échantillons, on peut s'intéresser aux variations d'un sous-échantillon à l'autre. Le test du C de Cochran et les statistiques h et k de Mandel font partie des méthodes adaptées à ce type d'études.

Définitions

Soit un échantillon $x_{11}, x_{12}, \ldots, x_{1n_1}, x_{12}, x_{22}, \ldots, x_{2n_2}, \ldots, x_{p1}, x_{p2}, \ldots, x_{pn_p}$, de données que l'on distingue pour leur appartenance à p groupes (par exemple des laboratoires) d'effectif respectif n_i $(i = 1, \ldots p)$. Soit $\overline{x}_i$ la moyenne du groupe i et s_i^2 la variance estimée pour le groupe i. On a :

$$\overline{x}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij}$$

et

$$s_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n} (x_{ij} - \overline{x}_i)^2$$

On suppose ici que les observations sont identiquement distribuées suivant une loi normale.

Statistique h de Mandel

La statistique h_i pour le groupe i, $(i = 1, \ldots, p)$ est donnée par :

$$h_i = \frac{\overline{x}_i - \overline{\overline{x}}}{s}$$

avec

$$\overline{\overline{x}} = \frac{1}{p} \sum_{i=1}^{p} \overline{x}_i \quad \text{et } s = \sqrt{\frac{1}{p-1} \sum_{i=1}^{p} (\overline{x}_i - \overline{\overline{x}})^2}$$

XLSTAT fournit les statistiques h_i pour chacun des groupes. Afin d'identifier les groupes pour lesquels la moyenne est potentiellement anormale, on peut calculer des valeurs critiques et un intervalle de confiance pour un niveau de signification α donné autour de la statistique h (Wilrich, 2013). La valeur critique est donnée par :

$$h_{crit}(p, \alpha) = \frac{(p-1)t_{p-2,1-\alpha/2}}{\sqrt{p(p-2+t^2_{p-2,1-\alpha/2})}}$$

où t correspond au quantile de distribution de Student pour $1-\alpha/2$ et avec $p-2$ degrés de liberté.

L'intervalle de confiance autour de h_i à $100(1-\alpha)\%$ est donc donné par $\left]-h_{i,crit};h_{i,crit}\right[$. XLSTAT affiche la valeur critique sur le graphique des valeurs h_i si les n_i sont constants.

Statistique k de Mandel

La statistique k_i pour le groupe i $(i=1,\ldots,p)$ est donnée par :

$$k_i = \frac{s_i}{\tilde{s}}$$

avec

$$s_i = \sqrt{\frac{1}{n_i-1}\sum_{j=1}^{n_i}\left(x_{ij}-\overline{x}_i\right)^2} \text{ et } \tilde{s} = \sqrt{\frac{1}{p}\sum_{i=1}^{p}s_i^2}$$

XLSTAT fournit les statistiques k_i pour chacun des groupes. Afin d'identifier les groupes pour lesquels la variance est potentiellement anormale, si les groupes sont de même taille n, on peut calculer des valeurs critiques et un intervalle de confiance pour un niveau de signification α donné autour de la statistique k (Wilrich, 2013). La valeur critique est donnée par :

$$k_{crit}(n,\alpha) = \sqrt{p(1+(p-1)F^{-1}_{1-\alpha,(p-1)(n-1),(n-1)})}$$

où $F^{-1}_{1-\alpha,v_1,v_2}$ est la valeur de la fonction de répartition inverse de la loi de Fisher pour la probabilité $1-\alpha$ avec v_1 et v_2 degrés de liberté.

L'intervalle de confiance (unilatéral) autour de k_i à $100(1-\alpha)\%$ est donc donné par $[0;k_{i,crit}]$. XLSTAT affiche la valeur critique sur le graphique des valeurs k_i si les n_i sont constants.

Z-scores

Parmi les résultats affichés par XLSTAT pour vous aider à identifier les valeurs extrêmes figurent les z-scores, qui correspondent aux x_i standardisés :

$$z_i = \frac{x_i-\overline{x}}{s}(i=1,\ldots,n)$$

Le problème de ces scores est qu'une fois l'intervalle admissible fixé (typiquement -1.96 et 1.96 pour un intervalle à 95%), toute valeur qui se trouve en dehors est considérée comme suspecte. Hors on sait que si l'on a 100 valeurs, il est statistiquement normal d'en avoir 5 en dehors de cette intervalle. Par ailleurs, par construction le z-score le plus élevé vaut au maximum :

$$\max_{i=1\ldots n} z_i \leq \frac{n-1}{\sqrt{n}}$$

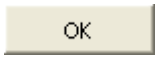
Iglewicz et Hoaglin (1993) recommandent l'utilisation d'un z-score modifié afin de mieux faire ressortir les valeurs extrêmes :

$$z_i = 0.6745\frac{x_i - \overline{x}}{MDA}(i = 1, \ldots, n)$$

où MDA est la médiane des déviations absolues (*Median Absolute Deviation*). L'intervalle acceptable est $]-3.5; 3.5[$ quelque soit n.

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas (mode colonnes), XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite (mode lignes), XLSTAT considère que les variables sont en lignes et les observations en colonnes.

Onglet **Général**:

Données : Si le format de données sélectionné est « une colonne par variable », sélectionnez les données pour les différentes variables. Si le format de données sélectionné est « une colonne par groupe », sélectionnez les échantillons associés aux différents groupes. Si le format de données sélectionné est « variances », sélectionnez les variances de chaque groupe. Si le format de données sélectionné est « moyennes », sélectionnez les moyennes de chaque groupe.

Identifiant de groupe / Taille de groupe : si le format de données sélectionné est « une colonne par variable », sélectionnez les données identifiant les groupes auxquels les données

sélectionnées correspondent. Si le format de données sélectionné est « Variances » ou "Moyennes", vous devez alors entrer la **taille de groupe** pour un plan équilibré.

Format des données : choisissez le format des données.

- **Une colonne/ligne par groupe** : activez cette option pour sélectionner une colonne (ou ligne en mode lignes) par groupe.
- **Une colonne/ligne par variable** : activez cette option pour que XLSTAT fasse autant de tests qu'il y a de colonnes (ou lignes en mode lignes), sachant que chaque colonne/ligne doit contenir le même nombre de lignes/colonnes, et qu'un identifiant de groupe doit par ailleurs être sélectionné.
- **Variances** : activez cette option si vos données correspondent à des variances déjà calculées. Vous devez alors entrer la taille de groupe (plan équilibré).
- **Moyennes** : activez cette option si vos données correspondent à des moyennes déjà calculées. Vous devez alors entrer la taille de groupe (plan équilibré).

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des colonnes : activez cette option si la première ligne des données sélectionnées contient un libellé.

Vous pouvez choisir le type de test à appliquer sur les données :

- **Statistique h de Mandel** : choisissez cet option pour calculer la statistique h de Mandel.
- **Statistique k de Mandel** : choisissez cet option pour calculer la statistique k de Mandel.

Onglet **Options**:

Hypothèse alternative : choisissez l'hypothèse alternative à utiliser pour le test.

Niveau de signification (%) : entrez le niveau de signification à utiliser pour les différents tests (valeur par défaut : 5%).

Onglet **Données manquantes** :

Ne pas accepter les valeurs manquantes : activez cette option pour que XLSTAT empêche la poursuite des calculs si des valeurs manquantes sont détectées.

Supprimer les observations : activez cette option pour supprimer les observations comportant des données manquantes.

Onglet **Sorties**:

Statistiques descriptives : activez cette option pour calculer et afficher les statistiques descriptives pour les différents échantillons.

Z-scores : activez cette option pour calculer et afficher les z-scores et le graphique correspondant. Vous avez le choix entre les **z-scores modifiés** ou les **z-scores classiques**. Dans le cas de ces derniers vous pouvez choisir deux limites à afficher sur les graphiques.

Résultats

Statistiques descriptives : dans ce tableau sont affichées les statistiques descriptives correspondant aux différents traitements.

Les résultats correspondants aux statistiques de Mandel sont ensuite affichés.

Les z-scores sont ensuite affichés s'ils ont été demandés.

Exemple

Un exemple de calcul des statistiques de Mandel est disponible sur leCentre d'aide XLSTAT :

http://www.xlstat.com/demo-mandelf.htm

Bibliographie

Barnett V. and Lewis T. (1980). Outliers in Statistical Data. John Wiley and Sons, Chichester, New York, Brisbane, Toronto.

Hawkins D.M. (1980). Identification of Outliers. Chapman and Hall, London.

Iglewicz B. and Hoaglin D. (1993). "Volume 16: How to Detect and Handle Outliers", The ASQC Basic References in Quality Control: Statistical Techniques, Edward F. Mykytka, Ph.D., Editor.

International Organization for Standardization (1994). ISO 5725-2: Accuracy (trueness and precision) of measurement methods and results—Part 2: Basic method for the determination of repeatability and reproducibility of a standard measurement method, Geneva.

Mandel J. (1991). The validation of measurement through interlaboratory studies. Chemometrics and Intelligent Laboratory Systems; **11**, 109-119.

Mandel J. (1985). A new analysis of interlaboratory test results. In: ASQC Quality Congress Transaction, Baltimore, 360-366.

Wilrich P. -T. (2013). Critical values of Mandel's h and k, the Grubbs and the Cochran test statistic. *Advances in Statistical Analysis*, 97(1), 1-10.

XLSTAT.ai

Easy Fit / Easy Predict

Utilisez la fonction Easy Fit afin de tester et comparer différents modèles prédictifs sur un même jeu de données. En fonction du type de variable à prédire (quantitative ou qualitative) et du type de variables explicatives, différents modèles prédictifs sont proposés. La fonction Easy Predict vous permet ensuite de prédire les valeurs de nouvelles observations sur les modèles précédemment générés.

Dans cette section :

Description

Introduction

Lorsque l'on cherche à prédire les valeurs d'une variable Y de nature quantitative, on parle de **régression**. Lorsque la variable Y à prédire est de nature qualitative, on parle alors de **classification**. XLSTAT possède plusieurs modèles d'apprentissage en régression et en classification. La fonction Easy Fit a été développée pour répondre principalement aux deux constats suivants :

- Lorsque l'on utilise un modèle prédictif au sein de XLSTAT, énormément de résultats sont disponibles (par défaut ou en option). Cela est indispensable pour certains experts mais cela peut aussi effrayer certains utilisateurs qui cherchent en premier lieu à savoir si le modèle exécuté donne de bons résultats et n'ont donc pas besoin de connaître tous les résultats.
- Actuellement, si on veut comparer plusieurs modèles sur un même jeu de données, il faut lancer différentes analyses de manière séparée tout en sélectionnant des échantillons de validation identiques. Ceci nécessite plusieurs étapes, parfois répétitives, qui peuvent prendre du temps.

Le but de la fonction Easy Fit est d'apporter des solutions aux deux points précédents. En effet, Easy Fit permet en fonction de la nature du problème (régression ou classification) de générer très rapidement plusieurs modèles différents sur un même jeu de données. Les résultats des différents modèles sont synthétisés afin de permettre à l'utilisateur de déterminer rapidement quel est le meilleur modèle.

La qualité des modèles est comparée grâce à différents indicateurs communs. Ces indicateurs sont calculés sur un échantillon de validation contenant 20% des observations sélectionnées aléatoirement.

Modèles de régression disponibles

Actuellement la fonction Easy Fit dispose des modèles de régression suivants (vous pouvez cliquer sur les différentes méthodes afin d'accéder au document d'aide associé) :

- Si les variables explicatives X sont uniquement **quantitatives** :
    - Régression linéaire
    - Forêts aléatoires de régression
    - K plus proches voisins
    - Machine à Vecteurs de Support
    - Régression LASSO
    - eXtreme Gradient Boosting
- Si les variables explicatives X sont uniquement **qualitatives** :
    - Analyse de la variance : ANOVA
    - Forêts aléatoires de régression
    - K plus proches voisins
    - Machine à Vecteurs de Support
    - Régression LASSO
    - eXtreme Gradient Boosting
- Si certaines variables explicatives X sont **quantitatives et d'autres qualitatives** :
    - Analyse de la covariance : ANCOVA
    - Forêts aléatoires de régression
    - K plus proches voisins
    - Machine à Vecteurs de Support
    - Régression LASSO
    - eXtreme Gradient Boosting

Modèles de classification disponibles

Actuellement la fonction Easy Fit dispose des modèles de classification suivants :

- Pour tout type de variables explicatives X, qu'elles soient uniquement **quantitatives**, uniquement **qualitatives** ou **quantitatives et qualitatives** :
    - Régression logistique
    - Forêts aléatoires de classification
    - K plus proches voisins
    - Machine à Vecteurs de Support
    - Analyse Factorielle Discriminante
    - eXtreme Gradient Boosting

Présentation des résultats

Au début de chaque feuille de résultats d'une analyse Easy Fit, vous trouverez un tableau récapitulatif.

Ce tableau contient plusieurs indicateurs de qualité des modèles effectués. Les indicateurs en régression sont les suivants : MAE, MCE, R^2, R^2 ajusté, AIC et SBC. En classification nous retrouvons les indicateurs suivants : Exactitude, Précision Sensibilité, Nombre de bien classés, Nombre de mal classés et F-score.

Ensuite, vous trouverez les résultats condensés de chaque analyse. Pour plus de détails sur ces résultats, veuillez vous reporter aux documents d'aide associés. Pour chaque modèle, vous trouverez au début des résultats deux boutons.

 : ce bouton vous permet d'ouvrir automatiquement la boîte de dialogue préremplie associée à la méthode complète. Vous aurez ainsi accès à toutes les options possibles. Cela peut être très utile si vous souhaitez obtenir plus de résultats sur la méthode en question.

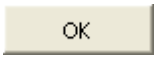
 : ce bouton vous permet d'ouvrir la boîte de dialogue de la fonction Easy Predict afin d'effectuer des prédictions sur de nouvelles observations. La fonction Easy Predict est présentée par la suite.

Easy Predict

Le but de Easy Fit est de trouver le meilleur modèle prédictif adapté aux données. Une fois que ce modèle a été trouvé, il est souvent nécessaire de prédire les valeurs de nouvelles observations n'ayant pas servi à l'apprentissage du modèle ni à sa validation, c'est le but de la fonction Easy Predict. En effet, en cliquant sur le deuxième bouton présenté au-dessus, vous ouvrirez une nouvelle boîte de dialogue vous permettant de sélectionner des nouvelles observations dont vous voulez prédire les valeurs à l'aide du modèle généré avec Easy Fit. Les résultats générés contiennent alors les prédictions associées à ces nouvelles observations.

Boîte de dialogue

La boîte de dialogue est composée de différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.

 : cliquez sur ce bouton pour lancer les calculs.

 : cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

 : cliquez sur ce bouton pour afficher l'aide.

 : cliquez sur ce bouton pour rétablir les options par défaut.

 : cliquez sur ce bouton pour effacer les sélections de données.

 : cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.

Type de la variable Y à prédire : choisissez le type de variable Y que vous voulez prédire. Si Y est quantitative alors des méthodes de régression vous seront proposées, si Y est qualitative alors des méthodes de classification vous seront proposées. Puis sélectionnez la variable à prédire dans le champ associé. Si le libellé de la variable a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

Types des variables explicatives X : choisissez le type des variables X explicatives (ou prédictives) à utiliser dans votre modèle. Vous pouvez utiliser uniquement des variables quantitatives, uniquement des variables qualitatives ou bien des variables quantitatives et des variables qualitatives. Ensuite vous pouvez sélectionner dans les champs associés vos variables explicatives. Si les libellés des variables ont été sélectionnés, veuillez vérifier que l'option « Libellés des variables » est activée.

Afficher les résultats dans :

- **Nouvelle feuille** : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif. Dans ce cas, vous avez la possibilité de donner un nom à la feuille de résultats. Si vous n'en spécifiez pas, un nom par défaut sera créé.
- **Nouveau classeur** : activez cette option pour afficher les résultats dans un nouveau classeur. Dans ce cas, vous avez la possibilité de donner un nom à la feuille de résultats. Si vous n'en spécifiez pas, un nom par défaut sera créé.
- **Cellule existante** : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Libellés des variables : activez cette option si la première ligne des données sélectionnées contient un libellé.

Sélectionnez les méthodes à utiliser : en fonction des types de variables utilisés, différentes méthodes sont proposées (voir la section Description pour la liste des différentes méthodes proposées).

Statistiques descriptives : activez cette option pour afficher des statistiques descriptives sur les différentes variables utilisées.

Résultats

Statistiques descriptives : les tableaux de statistiques descriptives présentent pour toutes les variables sélectionnées des statistiques simples. Pour les variables quantitatives sont affichés le nombre d'observations, le nombre de données manquantes, le nombre de données non manquantes, la moyenne, et l'écart-type (non biaisé). Pour les variables qualitatives sont affichés les noms des différentes modalités ainsi que leurs fréquences respectives.

Tableau de synthèse : ce tableau présente les coefficients d'ajustement jugeant de la qualité du modèle. Ces coefficients d'ajustement sont calculés sur l'échantillon de validation. Le nombre d'observations mal classées est un indicateur simple à utiliser pour comparer la performance de deux modèles de classification. Pour la régression, l'indicateur le plus souvent utilisé est la moyenne des carrés des erreurs. Pour ces deux indicateurs, le but est d'obtenir les valeurs les plus petites possibles.

Résultats synthétiques par méthode : pour chacune des méthodes réalisées les résultats les plus importants de la méthode sont affichés. Pour une meilleure compréhension des différents résultats vous pouvez accéder aux aides des différentes méthodes. Tous les liens vers les aides sont disponibles dans la section Description.

Exemple

Un exemple d'utilisation de la méthode Easy Fit et de la méthode Easy Predict est disponible sur le Centre d'aide XLSTAT à l'adresse suivante :

http://www.xlstat.com/demo-easyf.htm

Outils mathématiques

Calculateur de probabilités

Utilisez cet outil pour calculer, pour une loi de probabilité donnée, une densité de probabilité, la fonction de répartition, ou la fonction de répartition inverse.

Dans cette section :

Description

Une loi de probabilité $\mathbb{P}$ est un type particulier de fonction, que l'on appelle *mesure* en mathématique, qui permet de faire correspondre (grâce à ce que les mathématiciens appellent une *application*) à des événements (par exemple, la survenue d'un 2 lorsqu'on lance un dé), leur probabilité. Pour les mathématiciens on distingue l'univers Ω qui représente tous les possibles lorsque l'on fait une expérience aléatoire, de la tribu $\mathcal{A}$ qui regroupe les événements ayant la propriété d'être mesurables et qui est donc un sous-ensemble de l'univers. Dans le cas discret la tribu et l'univers sont identiques. Le triplet $(\Omega, \mathcal{A}, \mathbb{P})$ constitue un *espace mesurable*.

Un grand nombre de lois de probabilités a été développé afin de décrire des situations particulières où l'aléatoire intervient. L'aléatoire est une représentation de la méconnaissance ou de la connaissance imparfaite. Lorsque l'on jette un dé, une parfaite connaissance du mouvement de départ, de l'environnement, des forces en présence, ..., permettrait de savoir quand et dans quelle position le dé s'arrêterait. Néanmoins, cela est tellement complexe, que l'on préfère estimer que chaque face a une certaine probabilité d'être le résultat du jet et que la survenue d'un événement est aléatoire.

Une variable aléatoire, est également une fonction, qui fait correspondre à un événement un réel. Par exemple, on peut faire correspondre au jet d'une pièce, 0 si la pièce tombe sur pile et 1 si la pièce tombe sur face. Si cela est parfaitement arbitraire dans le cas d'une pièce, cela peut-être plus naturel si l'on compte le nombre de gens dans une queue au bureau de poste, ou si l'on mesure la température de l'air. Dans ce cas, on fera correspondre à l'événement "il y a 10 personnes dans la queue", le nombre 10, ou à "la température est de 19.8°C", le réel 19.8.

Dans le cas des variables discrètes, chaque événement a une probabilité non nulle dès lors qu'il est possible. Dans le cas des variables continues, chaque événement (possible) a une

probabilité non nulle de survenir, mais elle est si infime, que seule la loi de probabilité permet de la mesurer. Ainsi, par exemple, si on réalise une mesure de la température et que l'on sait que la mesure est entâchée d'erreur et que de ce fait la mesure suit une loi normale (la fameuse distribution avec une forme de cloche), l'événement "il fait 20°C", a une probabilité virtuellement nulle de survenir, alors qu'en revanche on saura donner la probabilité pour que l'on mesure une température comprise entre 19.5°C et 20.5°C. L'outil mathématique pour calculer cette probabilité s'appelle la **fonction de répartition**.

En théorie des probabilités, la loi de probabilité est d'abord décrite par sa fonction de répartition qui peut elle-même dépendre de différents paramètres. La loi la plus simple est la loi de Bernouilli où deux événements sont possibles (jet d'une pièce). Son paramètre est la probabilité p_0 pour que la pièce tombe sur pile et si la pièce n'est pas truquée, on a $p_0 = 0,5$. Pour la loi normale, loi essentielle en statistique, les paramètres sont la moyenne μ et la variance σ^2.

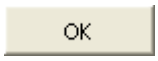
Pour une variable aléatoire continue, la fonction de répartition est une fonction croissante qui prend ses valeurs dans l'intervalle $[0;1]$.

Deux autres fonctions sont communément utilisées :

- La fonction de densité (ou densité de probabilité ou densité) qui est la fonction à intégrer pour calculer la fonction de répartition (cas de variables à densité). C'est le cas de toutes les lois proposées par XLSTAT.
- La fonction de répartition inverse : cette fonction permet, pour une probabilité donnée p, d'obtenir la valeur x de la variable aléatoire telle que la fonction de répartition en x vaille p.

Le calculateur de probabilité permet, pour toutes les lois proposées par XLSTAT, de calculer la fonction de densité, la fonction de répartition et la fonction de répartition inverse.

Boîte de dialogue



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.

Distribution : choisissez la loi de probabilité pour laquelle vous voulez effectuer des calculs. Vous pouvez ensuite entrer la valeur des différents paramètres de la loi. Une description des différentes lois de probabilité peut être trouvée ici.

Calculer : choisissez la fonction que vous voulez calculer (densité de probabilité, fonction de répartition ou fonction de répartition inverse), puis le point en lequel vous vous la calculer. Pour la fonction de répartition F, vous pouvez choisir entre :

- $< a$: le résultat retourné correspond à $F(a)$
- $> a$: le résultat retourné correspond à $1 - F(a)$
- $a << b$: le résultat retourné correspond à $(F(b) - F(a)$
- $< a \quad b <$: le résultat retourné correspond à $1 - (F(b) - F(a))$

Calculer : cliquez sur ce bouton pour afficher le résultat dans la zone *Résultats* de la boîte de dialogue.

Afficher les résultats dans :

- **Plage** : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.
- **Feuille** : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.
- **Classeur** : activez cette option pour afficher les résultats dans un nouveau classeur.

Afficher l'en-tête du rapport : désactivez cette option si vous souhaitez que le résultat soit affiché sur la première ligne de la feuille Excel (cas d'une sortie dans une feuille ou un classeur), et non après l'en-tête du rapport.

Résultats : les résultats sont affichés dans cette zone. Vous pouvez les copier (Ctrl C) pour les coller dans un autre logiciel.

Effacer : cliquez sur ce bouton pour effacer les résultats enregistrés dans la zone *Résultats* de la boîte de dialogue.

Résultats

Les résultats affichés par XLSTAT correspondent à la valeur de la fonction choisie au point choisi.

Exemple

Un exemple d'utilisation du calculateur de probabilités est disponible sur le Centre d'aide XLSTAT :

http://www.xlstat.com/demo-prcf.htm

Bibliographie

Saporta G. (1990). Probabilités, Analyse des Données et Statistique. Technip, Paris. 199-216.

Krishnamoorthy K. (2015). Handbook of Statistical Distributions with Applications. Chapman and Hall/CRC.

Opérations matricielles

Utilisez cet outil pour effectuer des opérations (addition, soustraction, produit) entre deux matrices.

Dans cette section :

Description

Cet outil vous permet d'effectuer des opérations entre deux matrices A et B. Les matrices peuvent être préalablement transformées :

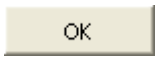
- la transposée peut être utilisée, elle est notée A' pour la matrice A;
- l'inverse peut être utilisée, elle est notée A^(-1).

Si une deuxième matrice B est sélectionnée alors différentes opérations matricielles sont disponibles :

- Addition : dans ce cas les matrices A et B doivent être de même taille (même nombre de lignes et même nombre de colonnes).
- Soustraction : dans ce cas les matrices A et B doivent être de même taille (même nombre de lignes et même nombre de colonnes).
- Produit : dans ce cas le nombre de lignes de la matrice B doit être égal au nombre de colonnes de la matrice A.

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.

 : cliquez sur ce bouton pour effacer les sélections de données.

 : cliquez sur ce bouton pour changer la façon dont vous chargez les données dans XLSTAT. Si le bouton possède une icône de souris, XLSTAT vous permet de faire une sélection à la souris des données. Si le bouton possède une icône de liste, XLSTAT vous permet de sélectionner les données à partir d'une liste.

Onglet **Général**:

Matrice A : sélectionnez les données correspondant à la matrice A. Si des en-têtes de colonnes ont été sélectionnés, veuillez vérifier que l'option « Libellés des variables » est activée.

- **Transposée : A'** : sélectionnez cette option si vous souhaitez utiliser la transposée de A à la place de A.
- **Inverse : A^(-1)** : sélectionnez cette option si vous souhaitez utiliser l'inverse de A à la place de A. Si la matrice n'est pas inversible, les calculs s'arrêteront.

Matrice B : sélectionnez cette option si vous voulez effectuer une opération entre deux matrices. Dans ce cas, sélectionnez les données correspondant à la matrice B. Si des en-têtes de colonnes ont été sélectionnés, veuillez vérifier que l'option « Libellés des variables » est activée.

- **Transposée : B'** : sélectionnez cette option si vous souhaitez utiliser la transposée de B à la place de B.
- **Inverse : B^(-1)** : sélectionnez cette option si vous souhaitez utiliser l'inverse de B à la place de B. Si la matrice n'est pas inversible, les calculs s'arrêteront.

Opération : sélectionnez ici l'opération souhaitée entre les matrices A et B : addition, soustraction ou produit (pour plus de détails, voir la partie description).

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des variables : activez cette option si la première ligne des données sélectionnées contient un libellé.

Afficher l'en-tête du rapport : désactivez cette option si vous souhaitez que le résultat soit affiché sur la première ligne de la feuille Excel (cas d'une sortie dans une feuille ou un classeur), et non après l'en-tête du rapport.

Résultat : en fonction des options choisies précédemment l'opération finale effectuée est affichée ici.

Onglet **Données manquantes** :

Ne pas accepter les valeurs manquantes : activez cette option pour que XLSTAT empêche la poursuite des calculs si des valeurs manquantes sont détectées.

Ignorer les données manquantes : si vous choisissez cette option, les données manquantes seront conservées. Toutes les opérations impliquant des données manquantes renverront des données manquantes.

Résultats

Résultat : la matrice que vous avez choisi de calculer est affichée.

Exemple

Un exemple d'utilisation de l'outil "opérations matricielles" est disponible sur le Centre d'aide XLSTAT à l'adresse suivante :

http://www.xlstat.com/demo-matf.htm

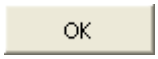
Outils

DataFlagger

Utiliser le DataFlagger pour faire ressortir des données qui sont comprises dans un intervalle ou en dehors d'un intervalle, ou qui sont égales à certaines valeurs.

Dans cette section :

Boîte de dialogue



: cliquez sur ce bouton pour marquer les données.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer de marquage.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT vous permet de sélectionner les données par colonnes ou par plage. Si la flèche est vers la droite, XLSTAT vous permet de sélectionner les données par lignes ou par plage.

Données : sélectionnez les données sur la feuille Excel.

Marquer une valeur ou un texte : activez cette option si vous souhaitez identifier et faire ressortir une valeur ou une série de valeurs dans la plage sélectionnée.

- **Valeur ou texte** : choisissez cette option pour rechercher et marquer une seule valeur ou une chaîne de caractères.
- **Liste de valeurs ou textes** : choisissez cette option pour rechercher et marquer une série de valeurs ou textes. Vous devez alors sélectionner dans une feuille Excel la série de valeurs ou textes en question.

Marquer un intervalle : activez cette option si vous souhaitez identifier et faire ressortir des valeurs comprises dans ou en dehors d'un intervalle. Définissez ensuite l'intervalle.

- **Dedans** : choisissez cette option pour rechercher et marquer les valeurs comprises dans un intervalle. Choisissez ensuite les types de bornes pour l'intervalle (ouvertes ou fermées), puis entrez la valeur des bornes.
- **Dehors** : choisissez cette option pour rechercher et marquer les valeurs comprises en dehors d'un intervalle. Choisissez ensuite les types de bornes pour l'intervalle (ouvertes ou fermées), puis entrez la valeur des bornes.

Police : utilisez les options suivantes pour modifier la police des valeurs correspondant aux règles de marquage.

- **Style** : choisissez le style de la police.
- **Taille** : choisissez la taille de la police.
- **Couleur** : choisissez la couleur de la police.

Cellule : utilisez l'option suivante pour modifier la couleur du fond de la cellule.

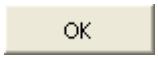
- **Couleur** : choisissez la couleur de la cellule.

Recherche du Min/Max

Utiliser cet outil pour repérer dans une plage de données les valeurs minimales et/ou maximales. Si la valeur minimale est rencontrée plusieurs fois, XLSTAT fait une sélection multiple des valeurs minimales vous permettant ensuite de naviguer de l'une à l'autre simplement en appuyant sur la touche « Entrée ».

Dans cette section :

Boîte de dialogue



: cliquez sur ce bouton pour lancer la recherche.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer de recherche.

: cliquez sur ce bouton pour afficher l'aide.

Données : sélectionnez les données sur la feuille Excel.

Trouver le minimum : activez cette option pour que XLSTAT recherche le ou les minimum dans la sélection. Si l'option « Sélection multiple » est activée et que plusieurs valeurs correspondant au minimum sont trouvées, elles seront toutes sélectionnées et vous pourrez naviguer de l'une à l'autre en cliquant sur la touche « Entrée » du clavier.

Trouver le maximum : activez cette option pour que XLSTAT recherche le ou les maximum dans la sélection. Si l'option « Sélection multiple » est activée et que plusieurs valeurs correspondant au maximum sont trouvées, elles seront toutes sélectionnées et vous pourrez naviguer de l'une à l'autre en cliquant sur la touche « Entrée » du clavier.

Sélection multiple : activez cette option pour que les différentes valeurs correspondant au minimum et/ou au maximum soient simultanément sélectionnées.

Supprimer les valeurs textuelles

Utilisez cet outil pour supprimer le contenu de cellules d'une feuille Excel qui contiennent des valeurs textuelles. Cet outil est particulièrement utile lorsque vous importez sous Excel des données numériques et que certaines données manquantes sont interprétées par Excel comme étant des chaînes vides.

Dans cette section :

Boîte de dialogue

OK : cliquez sur ce bouton pour supprimer les données textuelles.

Annuler : cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer de modification.

Aide : cliquez sur ce bouton pour afficher l'aide.

 : cliquez sur ce bouton pour rétablir les options par défaut.

 : cliquez sur ce bouton pour effacer les sélections de données.

Données : sélectionnez les données sur la feuille Excel.

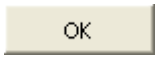
Nettoyer uniquement les cellules avec des chaînes vides : activez cette option pour que seules les cellules contenant des chaînes vides soient converties en cellules vides, sans format prédéfini.

Minuscules et Majuscules

Utilisez cet outil pour modifier la présence de majuscules ou minuscules dans des données textuelles.

Dans cette section :

Boîte de dialogue



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont vous chargez les données dans XLSTAT. Si le bouton possède une icône de souris, XLSTAT vous permet de faire une sélection à la souris des éléments. Si le bouton possède une icône de liste, XLSTAT vous permet de sélectionner les colonnes dénomminées par leur premier élément. Si le bouton possède une icône de feuille de papier de couleur orange, des boutons complémentaires avec un point d'interrogation  vont apparaître. Ceux-ci permettent de selectionner un fichier et les paramètres de lecture de celui-ci (voir Importer un fichier de données).

Onglet **Général** :

Données : sélectionnez les données sur la feuille Excel, dans une liste ou un fichier.

Minuscules : activez cette option pour convertir tous les textes en minuscules.

- **Première lettre en majuscule** : activez cette option pour que les mots aient leur première lettre en majuscule.
- **Premier mot uniquement** : activez cette option pour que seul le premier mot ait sa première lettre en majuscule.

Majuscules : activez cette option pour convertir tous les textes en majuscules.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des colonnes : activez cette option si la première ligne des données sélectionnées comprend un libellé qui ne doit pas être traité.

Afficher l'en-tête du rapport : désactivez cette option si vous souhaitez que le tableau des résultats commence dès la première ligne de la feuille Excel (cas d'une sortie dans une feuille ou un classeur), et non après l'en- tête du rapport.

Résultats

Les résultats sont affichés à l'endroit souhaité. Le tableau contient les chaînes traitées par les méthodes choisies.

Gestion des feuilles

Utilisez cet outil pour activer, afficher, cacher, ou supprimer une ou plusieurs feuilles contenues dans l'un des classeurs ouverts.

Dans cette section :

Boîte de dialogue

Lorsque vous lancez cet outil, une boîte de dialogue contenant la liste de toutes les feuilles (cachées ou non) de tous les classeurs est affichée.

Activer : cliquez sur ce bouton pour activer la première des feuilles sélectionnées.

Afficher : cliquez sur ce bouton pour afficher toutes les feuilles sélectionnées.

Cacher : cliquez sur ce bouton pour cacher toutes les feuilles sélectionnées.

Supprimer : cliquez sur ce bouton pour supprimer toutes les feuilles sélectionnées. Attention, la suppression des feuilles cachées est irréversible.

Annuler : cliquez sur ce bouton pour fermer la boîte de dialogue.

Aide : cliquez sur ce bouton pour afficher l'aide.

Supprimer les feuilles cachées

Utilisez cet outil pour supprimer les feuilles cachées générées par XLSTAT ou d'autres applications. XLSTAT génère des feuilles cachées pour créer certains graphiques. Cet outil permet de choisir les feuilles cachées à supprimer ou à garder.

Dans cette section :

Boîte de dialogue

Feuilles cachées : la liste des feuilles cachées est affichée. Sélectionnez les feuilles cachées que vous voulez supprimer.

Toutes : cliquez sur ce bouton pour sélectionner toutes les feuilles dans la liste.

Aucune : cliquez sur ce bouton pour désélectionner toutes les feuilles dans la liste.

Supprimer : cliquez sur ce bouton pour supprimer toutes les feuilles sélectionnées. Attention, la suppression des feuilles cachées est irréversible.

Annuler : cliquez sur ce bouton pour fermer la boîte de dialogue.

Aide : cliquez sur ce bouton pour afficher l'aide.

Afficher les feuilles cachées

Utilisez cet outil pour afficher les feuilles cachées générées par XLSTAT ou d'autres applications. XLSTAT génère des feuilles cachées pour créer certains graphiques. Cet outil permet de choisir les feuilles cachées à afficher.

Dans cette section :

Boîte de dialogue

Feuilles cachées : la liste des feuilles cachées est affichée. Sélectionnez les feuilles cachées que vous voulez afficher.

Toutes : cliquez sur ce bouton pour sélectionner toutes les feuilles dans la liste.

Aucune : cliquez sur ce bouton pour désélectionner toutes les feuilles dans la liste.

Afficher : cliquez sur ce bouton pour afficher toutes les feuilles sélectionnées.

Annuler : cliquez sur ce bouton pour fermer la boîte de dialogue.

Aide : cliquez sur ce bouton pour afficher l'aide.

Exporter vers GIF/JPG/PNG/TIF

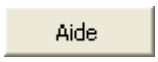
Utiliser cet outil pour exporter un graphique, une plage de données, un tableau, ou un objet quelconque vers un fichier graphique au format GIF, JPG, PNG ou TIF.

Dans cette section :

Boîte de dialogue

: cliquez sur ce bouton pour enregistrer l'objet sélectionné dans un fichier.

: cliquez sur ce bouton pour fermer la boîte de dialogue.



: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.

Format : choisissez le format graphique pour le fichier à générer.

Nom du fichier : entrez le nom du fichier à générer, ou choisissez le dans un répertoire donné.

Modifier la taille : activez cette option pour modifier la taille du graphique généré.

- **Largeur** : entrez la valeur en points de la largeur du graphique;
- **Hauteur** : entrez la valeur en points de la hauteur des graphique.

Afficher le quadrillage : activez cette option si vous souhaitez qu'en générant le graphique XLSTAT laisse figurer le quadrillage séparant les cellules. Cette option n'est active que lorsque des cellules ou des tableaux sont sélectionnés.

Ajouter des commentaires

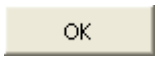
Utilisez cet outil pour créer ou ajouter des commentaires à des cellules de la feuille de calcul, en utilisant du contenu lui-même disponible dans des cellules.

Dans cette section :

Description

Les commentaires dans les cellules Excel sont intéressants car ils permettent de rendre optionnelle la visualisation de certaines informations, par exemple explicatives sur le contenu des cellules. Néanmoins leur chargement n'est pas forcément aisé. Grâce à cet outil de XLSTAT vous pourrez facilement créer ou modifier les commentaires.

Boîte de dialogue



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.

Commentaires : sélectionnez la ou les commentaires que vous souhaitez ajouter à des cellules. La disposition des commentaires doit correspondre à la disposition des cellules auxquelles vous voulez ajouter ces commentaires.

Cellules à commenter : sélectionnez la ou les cellules auxquelles vous voulez ajouter des commentaires. La disposition des commentaires doit correspondre à la disposition des cellules auxquelles vous voulez ajouter ces commentaires.

Fusionner : activez cette option pour que si une cellule est déjà commentée, le nouveau texte soit ajouté au commentaire existant.

Exemple

Un exemple d'ajout de commentaires est disponible à l'adresse suivante :

http://www.xlstat.com/demo-comf.htm

Analyse de données sensorielles

Cartographie externe des préférences (PREFMAP)

Utiliser cette méthode pour modéliser et représenter graphiquement les préférences de sujets pour une série d'objets en fonction de critères objectifs, ou de combinaisons linéaires de critères.

Dans cette section :

Description

La cartographie externe des préférences (en anglais *external preference mapping* - PREFMAP) permet de visualiser sur une même représentation graphique (en deux ou trois dimensions) d'une part des objets, et d'autre part des indications montrant le niveau de préférence de sujets (en général des consommateurs) en certains points de l'espace de représentation. Le niveau de préférence est représenté sur la carte de préférence sous formes de vecteurs, de points idéaux ou anti-idéaux, ou de courbes d'isopréférence en fonction du type de modèle choisi.

Les modèles sont eux-mêmes construits à partir de données objectives (par exemple des descripteurs physico-chimiques, ou des notes fournies par des experts sur des critères bien déterminés) ce qui permet d'interpréter la position des sujets et des produits en fonction des critères objectifs.

S'il n'y a que deux ou trois critères objectifs, les axes de l'espace de représentation sont définis par les critères eux-mêmes (éventuellement centrés-réduits pour éviter des effets d'échelle). En revanche, si le nombre de descripteurs est plus important, une méthode de réduction du nombre de dimensions doit être utilisée. En général, l'ACP est utilisée. Néanmoins, il est aussi possible d'utiliser l'analyse factorielle si l'on soupçonne l'existence de facteurs sous-jacents, ou un MDS (multidimensional scaling) si les données initiales sont des distances entre les produits. Si les descripteurs utilisés par les experts sont des variables qualitatives, on peut utiliser une ACM pour créer un espace à 2 ou trois dimensions.

Le PREFMAP peut être utilisé pour répondre aux questions suivantes :

Comment se positionne un produit par rapport à des produits concurrents ?

Quel est le produit concurrent le plus proche d'un produit donné ?

Quel type de consommateur préfère un produit ?

Pourquoi certains produits sont-ils préférés ?

Comment puis-je repositionner un produit pour qu'il soit encore davantage préféré par son cœur de cible ?

Quels nouveaux produits peuvent-ils être pertinents de créer ?

Modèles de préférence

Pour modéliser les préférences des sujets en fonction des critères objectifs ou de combinaison de critères objectifs (si une ACP a permis de générer l'espace à 2 ou 3 dimensions), quatre modèles ont été proposés dans le cadre du PREFMAP. Pour un sujet donné, si on désigne par y_i sa préférence pour le produit i, et par x_1, x_2, …, x_p les p critères ou combinaisons de critères (en général p = 2) décrivant le produit i, les modèles sont :

- Vectoriel : $y_i = a_0 + \sum_{j=1}^{p} a_j x_{ij}$
- Circulaire : $y_i = a_0 + \sum_{j=1}^{p} a_j x_{ij} + b \sum_{j=1}^{p} x_{ij}^2$
- Elliptique : $y_i = a_0 + \sum_{j=1}^{p} a_j x_{ij} + \sum_{j=1}^{p} b_j x_{ij}^2$
- Quadratique : $y_i = a_0 + \sum_{j=1}^{p} a_j x_{ij} + \sum_{j=1}^{p} b_j x_{ij}^2 + \sum_{j=1}^{p-1} \sum_{k=j+1}^{p} c_{jk} x_{ij} x_{ik}$

Les coefficients a, b, c sont estimés par régression linéaire multiple. On peut remarquer que les modèles sont classés du plus simple au plus complexe. XLSTAT permet, soit de choisir un modèle à utiliser pour tous les sujets, soit de retenir pour un sujet donné le modèle donnant le meilleur résultat au sens de la p-value du F de Fisher ou du test du F-ratio.

Le modèle **vectoriel** permet de représenter les individus sur la carte sensorielle sous forme de vecteurs. La taille des vecteurs est fonction du R² du modèle : plus le vecteur est long, meilleur est le modèle correspondant. La préférence du sujet sera d'autant plus forte que l'on sera loin dans la direction indiquée par le vecteur. L'interprétation de la préférence peut se faire en projetant sur les vecteurs les différents produits (préférence produit). L'inconvénient du modèle vectoriel est qu'il néglige le fait que pour certains critères (le salé ou la température par exemple), on peut avoir une croissance de la préférence jusqu'à un optimum puis une décroissance.

Le modèle **circulaire** permet de prendre en compte cette notion d'optimum. Si la surface correspondant au modèle a un maximum en terme de préférence (cela se produit si le coefficient b estimé est négatif), on parle de point idéal (venant de l'anglais *ideal point* à comprendre comme « point correspondant à l'idéal »). Si la surface a au contraire un minimum (cela se produit si le coefficient b estimé est positif), on parle de point anti-idéal (venant de l'anglais *anti-ideal point* à comprendre comme « point correspondant à l'opposé de l'idéal »). Avec le modèle circulaire, on peut tracer des lignes circulaires d'isopréférence autour du point idéal ou anti- idéal.

Le modèle **elliptique** est proche du modèle circulaire. Plus souple, il permet de mieux tenir compte d'effets d'échelle. L'inconvénient de ce modèle est que l'optimum du modèle n'existe pas toujours : comme avec le modèle circulaire, on peut obtenir un point idéal, ou un point anti-idéal, mais il arrive aussi que l'on obtienne un point selle (de la forme de la surface, rappelant une selle de cheval) si tous les coefficients b_j ne sont pas du même signe. Le point selle n'est pas facilement interprétable. Il correspond uniquement à une zone ou la préférence est moins sensible aux variations.

Enfin, le modèle **quadratique** permet de modéliser des structures de préférence plus complexes, en tenant notamment compte d'interactions. Comme avec le modèle elliptique, on peut obtenir un point idéal, un point anti- idéal, ou un point selle si tous les coefficients b_j ne sont pas du même signe.

Carte des préférences

La carte des préférences est une vision synthétique de trois types d'éléments :

les sujets (ou groupes de sujets si une classification des sujets a d'abord été effectuée) représentés au travers du modèle correspondant par un vecteur, un point idéal (noté +), un point anti-idéal (noté -), ou un point selle (noté o) ;

les objets dont la position sur la carte est déterminée par leurs coordonnées ;

les descripteurs, qui correspondent aux axes de représentation, ou leur sont liés (lorsqu'une ACP précède le PREFMAP, on étudiera le biplot issu de l'ACP pour interpréter la position des objets en fonction des critères objectifs).

Le PREFMAP, avec l'interprétation qu'en permet la carte des préférences, est un outil d'aide à l'interprétation et à la décision potentiellement très puissant puisqu'il permet de relier des données de préférence à des données objectives. Cependant, il faut que les modèles associés aux sujets soient bien ajustés pour que l'interprétation soit fiable.

Scores de préférence

Le score de préférence de chaque objet pour un sujet donné, dont la valeur est comprise entre 0 (minimum) et 1 (maximum), est calculé à partir de la prédiction du modèle correspondant au sujet. Le score est d'autant plus élevé que le produit est préféré. Des scores de préférence des différents produits, on déduit un ordre de préférence des objets, pour chacun des sujets.

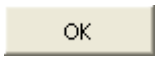
Contour plot

Le contour plot (courbes de niveau) permet de visualiser, sur un graphique dont les axes sont les mêmes que ceux de la carte des préférences, les régions correspondant à différents niveaux de consensus de préférence. En chaque point du graphique, on calcule le pourcentage de sujets pour lesquels la préférence calculée à partir du modèle est supérieure à leur préférence moyenne. Dans les régions correspondant aux couleurs froides (bleus), une faible proportion de modèles donne des préférences élevées. Au contraire, dans les régions

correspondant aux couleurs chaudes (rouge), une forte proportion de modèles donne des préférences élevées.

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.

Onglet **Général** :

Y / Données de préférence : sélectionnez les données de préférence. Le tableau doit contenir en ligne les différents objets (produits) étudiés, et en colonne les sujets (en mode transposé, cela doit être le contraire). Si des en- têtes de colonnes ont été sélectionnés, veuillez vérifier que l'option « Libellés des variables » est activée.

Remarque : XLSTAT considère que les préférences sont des données croissantes (plus un sujet apprécie un objet, plus la préférence est élevée).

Centrer : activez cette option si vous voulez centrer les données de préférence avant de commencer les calculs.

Réduire : activez cette option si vous voulez réduire les données de préférence avant de commencer les calculs.

X / Configuration : sélectionnez les données qui correspondent aux descripteurs objectifs ou à une configuration en deux ou trois dimensions si une méthode a déjà été utilisée pour générer la configuration. Si des en-têtes de colonnes ont été sélectionnés, veuillez vérifier que l'option « Libellés des variables » est activée.

Transformation préliminaire : activez cette option si vous souhaitez transformer les données.

- **Normalisation** : activez cette option pour centrer-réduire les données de la configuration X avant de réaliser le PREFMAP.
- **ACP (Pearson)** : activez cette option pour que XLSTAT transforme les descripteurs sélectionnés au moyen d'une Analyse en Composantes Principales (ACP) normée. Le nombre de composantes utilisées pour la suite des calculs est déterminé par le nombre de **dimensions** choisies.
- **ACP (Covariance)** : activez cette option pour que XLSTAT transforme les descripteurs sélectionnés au moyen d'une Analyse en Composantes Principales (ACP) non normée. Le nombre de composantes utilisées pour la suite des calculs est déterminé par le nombre de **dimensions** choisies.
- **PLS(Std) / PLS** : activez l'une de ces options pour que XLSTAT transforme les descripteurs sélectionnés en extrayant les componsantes PLS. Celles-ci présentent l'avantage de prendre en compte non seulement la structure de covariance entre le X, mais aussi celle entre les Y et entre les X et les Y. L'option (Std) a pour effet qu'une standardisation des Y est appliquée. Le nombre de composantes utilisées pour la suite des calculs est déterminé par le nombre de **dimensions** choisies.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des variables : activez cette option si la première ligne des données sélectionnées (variables dépendantes et explicatives, libellés des observations) contient un libellé.

Libellés des objets : activez cette option si vous voulez utiliser des libellés d'objets pour l'affichage des résultats. Si l'option « Libellés des variables » est activée, la première cellule de la sélection doit comprendre un en-tête. Si vous n'activez pas cette option, des libellés seront automatiquement créés (Obs1, Obs2, …).

Onglet **Options** :

Modèle : choisissez le type de modèle à utiliser pour relier les préférences à la configuration X si l'option « Rechercher le meilleur modèle » n'a pas été activée (voir onglet Options).

Dimensions : entrez le nombre de dimensions à utiliser pour le modèle PREFMAP (valeur par défaut : 2).

Rechercher le meilleur modèle : activez cette option afin de permettre à XLSTAT de trouver pour chaque sujet quel est le modèle le plus performant.

- **F-ratio** : activez cette option pour sélectionner le modèle donnant la meilleure p-value associée au F-ratio. Un modèle plus complexe est retenu si la p-value associée au F-ratio est inférieure au seuil de signification choisi ci-dessous.
- **F** : activez cette option pour sélectionner le modèle donnant la meilleure p-value associée au F de Fisher.

Niveau de signification (%) : entrez le niveau de signification. Les p-values des modèles sont affichées en gras lorsqu'elles sont inférieures à ce niveau.

Poids : si vous voulez attribuer des poids aux différents sujets (par exemple, parce qu'en réalité ce sont des groupes de sujets), vous pouvez activer cette option et sélectionner les poids correspondants.

Ces options ne sont visibles que dans le cas où une transformation préliminaire par ACP a été demandée.

Variables supplémentaires : activez cette option si vous voulez calculer les coordonnées a posteriori pour des variables qui ne sont pas prises en compte pour le calcul des axes factoriels (variables passives, par opposition aux variables actives).

- **Quantitatives** : activez cette option si vous disposez de variables quantitatives supplémentaires. Si des en-têtes de colonnes ont été sélectionnés pour le tableau principal, veillez à ce qu'un libellé soit aussi présent pour les variables de cette sélection.

Onglet **Prédiction** :

Cet onglet n'est pas visible si une transformation préliminaire par ACP a été demandée.

Prédiction : activez cette option si vous souhaitez sélectionner des données à utiliser en mode prédiction. Si vous activez cette option, vous devez veiller à ce que les données de prédiction soient organisées comme les données d'estimation : mêmes variables, même ordre dans les sélections. En revanche, vous ne devez pas sélectionner de libellés de variables : la première ligne des sélections décrites ci-dessous doit être une ligne de données.

X / Configuration : activez cette option pour sélectionner les données de la configuration à utiliser pour des prédictions. La première ligne ne doit pas comprendre d'en-tête.

Libellés des objets : activez cette option, si vous voulez utiliser des libellés d'objets pour les données de prédiction. La première ligne ne doit pas comprendre d'en-tête. Si vous n'activez pas cette option, des libellés seront automatiquement créés (PredObs1, PredObs2, …).

Libellés des variables : activez cette option si la première ligne des données sélectionnées (variables explicatives, libellés des observations) contient un libellé.

Onglet **Données manquantes** :

Supprimer les observations : activez cette option pour supprimer les observations comportant des données manquantes.

Estimer les données manquantes : activez cette option pour estimer les données manquantes avant le début des calculs.

- **Moyenne ou mode** : activez cette option pour estimer les données manquantes en utilisant la moyenne (variables quantitatives) ou le mode (variables qualitatives) pour les variables correspondantes.
- **Plus proche voisin** : activez cette option pour estimer les données manquantes d'une observation en recherchant le plus proche voisin de l'observation.

Onglet **Sorties** :

Statistiques descriptives : activez cette option pour afficher les statistiques descriptives pour les variables sélectionnées.

Corrélations : activez cette option pour afficher la matrice de corrélation pour les différentes variables sélectionnées.

Analyse de la variance : activez cette option pour afficher le tableau d'analyse de la variance pour les différents modèles.

Coefficients des modèles : activez cette option pour afficher les paramètres des modèles.

Prédictions des modèles : activez cette option pour afficher les prédictions des modèles.

Scores de préférence : activez cette option pour afficher les scores de préférence sur une échelle de 0 à 1.

Rangs des scores de préférence : activez cette option pour afficher les rangs correspondant aux scores de préférence.

Objets classés : activez cette option pour afficher les objets dans l'ordre décroissant de préférence pour chacun des sujets.

Points idéaux potentiels : activez cette option pour afficher le tableau des points idéaux potentiels.

Zone d'admission : activez cette option pour afficher la zone d'admission pour les points idéaux. Cette zone est définie par un minimum et un maximum. En dehors de cette zone, il est très peu vraisemblable qu'il y ait des points idéaux.

Dans le cas où une transformation préliminaire par ACP a été demandée, les options suivantes sont disponibles :

Valeurs propres : activez cette option pour afficher les valeurs propres de l'ACP.

Coordonnées des variables : activez cette option pour afficher les coordonnées des variables (*factor loadings* en anglais). Les coordonnées sont égales aux corrélations entre les composantes principales et les variables d'origine dans le cas d'une ACP normée.

Corrélations Composantes/Variables : activez cette option pour afficher les corrélations entre les composantes principales et les variables d'origine.

Coordonnées des observations : activez cette option pour afficher les coordonnées des observations (*factor scores* en anglais) dans le nouvel espace créé par l'ACP. Ces coordonnées sont ensuite utilisées pour le PREFMAP.

Onglet **Graphiques (ACP)** :

Cet onglet n'est visible que dans le cas où une transformation préliminaire par ACP a été demandée.

Graphiques de corrélations : activez cette option pour afficher les graphiques mettant en jeu des corrélations entre des composantes et des variables initiales.

- **Vecteurs** : activez cette option pour afficher les variables d'origine sous forme de vecteurs.

Graphiques des observations : activez cette option pour afficher les graphiques de représentation des observations dans le nouvel espace.

- **Etiquettes** : activez cette option pour afficher les étiquettes des observations sur les graphiques. Le nombre d'étiquettes affichées peut être modulé à l'aide de l'option de filtrage.

Biplots : activez cette option pour afficher les graphiques de représentation simultanée des observations et des variables d'origine dans le nouvel espace.

- **Vecteurs** : activez cette option pour afficher les variables d'origine sous forme de vecteurs.
- **Etiquettes** : activez cette option pour afficher les étiquettes des observations sur les biplots. Le nombre d'étiquettes affichées peut être modulé à l'aide de l'option de filtrage.

Type de biplots : choisissez le type de biplot que vous souhaitez afficher. Voir la section description de l'ACP pour plus de détails.

- **Biplot de corrélation** : activez cette option pour afficher des biplots de corrélation.
- **Biplot de distance** : activez cette option pour afficher des biplots de distance.
- **Biplot symétrique** : activez cette option pour afficher des biplots symétriques.
- **Coefficient** : choisissez le coefficient dont la racine carrée sera multipliée par les coordonnées des variables. Ce coefficient vous permettra d'ajuster la position des points variables dans le biplot afin de rendre ce dernier plus lisible. Si ce coefficient est différent de 1, la longueur des vecteurs variables n'est plus interprétable en termes d'écart-type (biplot de corrélation) ou de contribution (biplot de distance).

Etiquettes colorées : activez cette option pour que les étiquettes soient de la même couleur que les points correspondants.

Onglet **Graphiques** :

Carte des préférences : activez cette option pour afficher la carte des préférences.

- **Afficher les points idéaux** : activez cette option pour afficher les points idéaux.
- **Afficher les points anti-idéaux** : activez cette option pour afficher les points anti-idéaux.
- **Afficher les points selle** : activez cette option pour afficher les points selle.
- **Restriction du domaine** : activez cette option pour n'afficher les points solution (idéaux, anti-idéaux, selle) que s'ils se trouvent à l'intérieur d'un domaine à définir. Entrez alors la taille de zone à utiliser pour l'affichage : cette taille est exprimée en % de la zone délimitée par la configuration X (valeur comprise entre 100 et 500).
- **Longueur des vecteurs** : les options ci-dessous permettent de déterminer la longueur des vecteurs sur la carte de préférence, lorsqu'un modèle vectoriel est utilisé.
- **Coefficients** : choisissez cette option pour que la longueur des vecteurs soit uniquement déterminée par les coefficients du modèle vectoriel.
- **R²** : choisissez cette option pour que la longueur des vecteurs soit déterminée par la valeur du R² du modèle. Ainsi, mieux un modèle est ajusté, plus long est le vecteur correspondant sur la carte.
- **=** : choisissez cette option pour que tous les vecteurs soient de la même taille.
- **Facteur d'allongement** : utilisez cette option pour multiplier la longueur de tous les vecteurs par une valeur arbitraire (valeur par défaut :1)

Modèle circulaire :

- **Afficher des cercles** : entrez le nombre de cercles d'isopréférence à afficher.

Courbes de niveau : activez cette option pour afficher le contour plot (voir description). Vous pouvez alors choisir entre les options suivantes :

- **Seuil / Moyenne (%)** : entrez le niveau par rapport à la moyenne des préférences tous sujets confondus, exprimé en %, à partir duquel on peut considérer qu'un sujet a une préférence pour un produit (la valeur par défaut, 100, correspond à la moyenne).
- **Seuil (Valeur)** : entrez le niveau absolu à partir duquel on peut considérer qu'un sujet a une préférence pour un produit.
- **Echelle de couleur** : choisissez vos couleurs.

PREFMAP & Courbes de niveau : activez cette option pour afficher la superposition de la carte de préférence et du contour plot. Trois niveaux de qualité sont proposés. Si vous observez des petits défauts dans le graphique, vous pouvez augmenter le nombre de points calculés.

Résultats

Statistiques simples : dans ce tableau sont affichés pour tous les sujets et toutes les dimensions de la configuration X (avant transformation si une transformation a été demandée), le nombre de données non manquantes, la moyenne, et l'écart-type (non biaisé).

Matrice de corrélation : ce tableau est affiché afin de vous permettre d'avoir un aperçu des corrélations entre les différentes variables sélectionnées.

Sélection du modèle : ce tableau permet de visualiser quel modèle a été utilisé pour chacun des sujets. Si le modèle n'est pas un modèle vectoriel, le type de point solution est affiché (idéal, anti-idéal, selle) avec ses coordonnées.

Analyse de la variance : dans ce tableau sont affichées les statistiques permettant d'évaluer la qualité de l'ajustement du modèle (R^2, F, et Pr>F). Lorsque la p-value (Pr>F) est inférieure au niveau de signification choisi, elle est affichée en gras. Si l'option F-ratio a été choisie dans l'onglet « Options », les résultats du test du F-ratio sont affichés (valeur du F et p-value associée).

Coefficients du modèle : dans ce tableau sont affichés, pour chaque sujet, les différents coefficients du modèle retenu.

Prédictions du modèle : ce tableau correspond aux préférences estimées par le modèle pour chaque sujet et chaque produit. Remarque : si les préférences ont été centrées-réduites, ces résultats correspondent aussi à des préférences centrées-réduites.

Scores de préférence de 0 à 1 : ce tableau correspond aux prédictions remises sur une échelle de 0 à 1.

Rangs des scores de préférence : dans ce tableau sont affichés les rangs des scores de préférence. Plus le rang est élevé, plus la préférence est élevée.

Objets classés par ordre croissant de préférence : dans ce tableau sont affichés par ordre croissant de préférence, pour chaque sujet, la liste des objets. Autrement dit, la dernière ligne correspond aux objets préférés des sujets, selon les modèles de préférence.

Pourcentage de sujets satisfaits : dans ce tableau sont affichés pour chaque produit le pourcentage de sujets étant au-dessus du seuil fixé.

La **carte des préférences** et le **contour plot** sont ensuite affichés. Sur la carte de préférence, les points idéaux sont figurés par (+), les points anti-idéaux par (-) et les points selle par (o).

Si l'option correspondante a été activée et si vous utilisez Excel 2003 ou supérieure, vous pouvez visualiser la **superposition de la carte des préférences et du contour plot**. Ce graphique peut être redimensionné, mais pour que la superposition soit maintenue après le redimensionnement, vous devez cliquer dans la feuille Excel puis à nouveau sur le graphique.

Les **points idéaux potentiels** et la **zone d'admission** permettent d'orienter les équipes de recherche et développement en leur indiquant des points idéaux potentiels (le nombre de sujets pour lesquels le modèle associé donne une préférence au-dessus du critère choisi est maximal) et une zone en dehors de laquelle il est au contraire peu probable qu'il y ait un point idéal.

Exemple

Un exemple de Preference Mapping est disponible sur le Centre d'aide XLSTAT :

http://www.xlstat.com/demo-prefmapf.htm

Bibliographie

Danzart M. and Heyd B. (1996). Le modèle quadratique en cartographie des préférences. 3ème Congrès Sensometrics, ENITIAA.

Naes T. and Risvik E. (1996). Multivariate Analysis of Data in Sensory Science. Elsevier Science, Amsterdam.

Schlich P. and McEwan J.A. (1992). Cartographie des préférences. Un outil statistique pour l'industrie agro-alimentaire. *Sciences des aliments*, **12**, 339-355.

Cartographie interne des préférences

Utiliser la cartographie interne des préférences (CIP) pour analyser les notes attribuées à les P produits à J juges (consommateurs, experts, ...): Alors que la cartographie externe des préférences (PREFMAP) permet de relier les notes données par des consommateurs à des données sensorielles (mesures chimiques, évaluations fournies par des experts), pour la cartographie interne seules les données de préférences sont nécessaires. La CIP est basée sur l'ACP et comprend deux options pour améliorer la visualisation des résultats.

Dans cette section :

Description

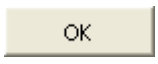
La cartographie interne (CIP) est basée sur l'Analyse en Composantes Principales (ACP) pour permettre d'identifier les produits qui correspondent aux attentes de groupes de consommateurs.

Pour plus d'informations sur l'ACP, vous pouvez lire la description disponible dans la section dédiée à cette méthode. Alors que l'ACP ne filtre pas les variables, cet outil permet d'éliminer a posteriori les juges qui ne sont pas assez bien affichés sur une carte à 2 dimensions. La mesure permettant d'évaluer la qualité de la projection d'un espace à d dimensions vers un espace plus petit d'un point est nommée communalité. Il peut aussi être interprété comme la somme des cosinus carrés entre le vecteur et les axes du sous-espace.

Le biplot qui est généré ici n'est pas un vrai biplot en ce sens que tous les juges retenus sont déplacés sur un cercle virtuel entourant les points produits en vue de faciliter l'interprétation visuelle.

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.

Onglet **Général**:

Le champ principal de saisie des données vous permet de sélectionner alternativement trois types de tableaux :

Tableau produits\juges : sélectionnez un tableau comprenant N produits notés par P juges. Si des en-têtes de colonnes ont été sélectionnés, veuillez vérifier que l'option « Libellés des variables » est activée.

Type d'ACP : choisissez entre corrélation (ACP normée), covariance (ACP non normée) et Spearman pour effectuer l'ACP sur une matrice de corrélation de Spearman.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des variables : activez cette option si la première ligne des données sélectionnées (Tableau produits\juges, libellés des produits, poids) contient un libellé.

Libellés des produits : activez cette option si vous voulez utiliser des libellés d'observations pour l'affichage des résultats. Si l'option « Libellés des juges » est activée, la première cellule de la sélection doit comprendre un en-tête. Si vous n'activez pas cette option, des libellés seront automatiquement créés (Obs1, Obs2, …).

Poids : activez cette option si vous voulez pondérer les observations. Si vous n'activez pas cette option, les poids seront tous considérés comme valant 1. Les poids doivent être impérativement supérieurs ou égaux à 0. Si un en-tête de colonne a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

Onglet **Options**:

Filtrer les facteurs : vous pouvez activer l'une des deux options suivantes afin de réduire le nombre de facteurs pour lesquels les résultats sont affichés :

- **% minimum** : activez cette option puis saisissez le pourcentage minimum de la variabilité totale que doivent représenter les facteurs retenus.
- **Nombre maximum** : activez cette option pour fixer le nombre maximum de facteurs à prendre en compte.

Normalisation : choisissez comment sont calculées les corrélations (ou covariance) : dénominateur (n) ou (n - 1).

Rotation : activez cette option si vous voulez appliquer une rotation à la matrice des coordonnées factorielles.

- **Nombre de facteurs** : entrez le nombre de facteurs pour lesquels la rotation sera appliquée.
- **Méthode** : choisissez la méthode de rotation à utiliser. Pour certaines méthode la valeur d'un paramètre doit être entrée (Kappa pour Orthomax, Tau pour Oblimin, et la puissance pour Promax).
- **Normalisation de Kaiser** : activez cette option pour appliquer la normalisation de Kaiser pendant le calcul des rotations.

Onglet **Données supplémentaires** :

Observations supplémentaires : activez cette option si vous voulez calculer les coordonnées et représenter des individus supplémentaires. Ces individus ne sont pas pris en compte pour le calcul des axes factoriels (observations passives, par opposition à observations actives). Si des libellés de variables sont présents pour les observations supplémentaires vous devez activer l'option « Libellés des variables pour les obs. ».

Variables supplémentaires : activez cette option si vous voulez calculer les coordonnées a posteriori pour des variables qui ne sont pas prises en compte pour le calcul des axes factoriels (variables passives, par opposition aux variables actives).

- **Quantitatives** : activez cette option si vous disposez de variables quantitatives supplémentaires. Si des en-têtes de colonnes ont été sélectionnés pour le tableau principal, veillez à ce qu'un libellé soit aussi présent pour les variables de cette sélection.
- **Qualitatives** : activez cette option si vous disposez de variables qualitatives supplémentaires. Si des en-têtes de colonnes ont été sélectionnés pour le tableau principal, veillez à ce qu'un libellé soit aussi présent pour les variables de cette sélection.
- **Colorer les observations** : activez cette option pour que les observations soient affichées avec des couleurs différentes selon la valeur de la première variable qualitative

supplémentaire.

- **Afficher les barycentres** : activez cette option pour afficher les barycentres correspondant aux modalités des différentes variables qualitatives supplémentaires sélectionnées sur les graphiques des observations.

Onglet **Données manquantes** :

Ne pas accepter les valeurs manquantes : activez cette option pour que XLSTAT empêche la poursuite des calculs si des valeurs manquantes sont détectées.

Supprimer les observations : activez cette option pour supprimer les observations comportant des données manquantes.

Suppression par paires : activez cette option pour supprimer les observations comportant des données manquantes uniquement lorsque les variables impliquées dans les calculs comportent des données manquantes. Par exemple lors du calcul d'une corrélation entre deux variables, une observation ne sera ignorée que si la donnée correspondant à l'une des deux variables est manquante.

Estimer les données manquantes : activez cette option pour estimer les données manquantes avant le début des calculs.

- **Moyenne ou mode** : activez cette option pour estimer les données manquantes en utilisant la moyenne (variables quantitatives) ou le mode (variables qualitatives) pour les variables correspondantes.
- **Plus proche voisin** : activez cette option pour estimer les données manquantes d'une observation en recherchant le plus proche voisin de l'observation.

Onglet **Sorties**:

Statistiques descriptives : activez cette option pour afficher les statistiques descriptives pour les variables sélectionnées.

Corrélations : activez cette option pour afficher la matrice de corrélation ou de covariance en fonction du type d'options choisi dans l'onglet « Général ».

- **Tester la significativité** : dans le cas où une corrélation a été choisie dans l'ongle « Général » de la boîte de dialogue, activez cette option pour tester la significativité des corrélations.
- **Test de sphéricité de Bartlett** : activez cette option pour effectuer le test de sphéricité de Bartlett.
- **Niveau de signification (%)** : entrez le niveau de signification pour les tests ci-dessus.
- **Kaiser-Meyer-Olkin** : activez cette option pour calculer la statistique de la précision d'échantillonnage (*Measure of Sampling Adequacy* en anglais) de Kaiser-Meyer-Olkin.

Valeurs propres : activez cette option pour afficher le tableau et le graphique (*scree plot*) des valeurs propres.

Coordonnées des variables : activez cette option pour afficher les coordonnées des variables dans l'espace des facteurs (*factor loadings* en anglais).

Corrélations Variables/Facteurs : activez cette option pour afficher les corrélations entre les facteurs et les variables.

Coordonnées des observations : activez cette option pour afficher les coordonnées des observations (*factor scores* en anglais) dans le nouvel espace créé par l'ACP.

Contributions : activez cette option pour afficher les tableaux des contributions pour les variables et les observations.

Cosinus carrés : activez cette option pour afficher les tableaux des cosinus carrés pour les variables et les observations.

Filtrer les juges : activez cette option pour éliminer des différentes sorties les juges pour lesquels la communalité est inférieure à un seuil donné.

Onglet **Graphiques** :

Graphiques de corrélations : activez cette option pour afficher les graphiques mettant en jeu des corrélations entre des composantes et des variables initiales.

- **Vecteurs** : activez cette option pour afficher les variables d'origine sous forme de vecteurs.

Graphiques des observations : activez cette option pour afficher les graphiques de représentation des observations dans le nouvel espace.

- **Etiquettes** : activez cette option pour afficher les étiquettes des observations sur les graphiques. Le nombre d'étiquettes affichées peut être modulé à l'aide de l'option de filtrage.

Biplots : activez cette option pour afficher les graphiques de représentation simultanée des observations et des variables d'origine dans le nouvel espace.

- **Vecteurs** : activez cette option pour afficher les variables d'origine sous forme de vecteurs.
- **Etiquettes** : activez cette option pour afficher les étiquettes des observations sur les biplots. Le nombre d'étiquettes affichées peut être modulé à l'aide de l'option de filtrage.
- **Déplacer vers le cercle** : activez cette option pour déplacer tous les points correspondant aux juges vers un cercle entourant tous les points produits.

Résultats

Statistiques descriptives : le tableau de statistiques descriptives présente pour toutes les variables sélectionnées des statistiques simples. Sont affichés le nombre d'observations, le

nombre de données manquantes, le nombre de données non manquantes, la moyenne, et l'écart-type (non biaisé).

Matrice de corrélation/de covariance : ce tableau correspond aux données qui sont ensuite utilisées pour les calculs. Le type de corrélation dépend de l'option qui a été choisie dans l'onglet « Général » de la boîte de dialogue. Dans le cas de corrélations, les corrélations significatives sont affichées en gras.

Test de sphéricité de Bartlett : les résultats du test de sphéricité de Bartlett sont affichés. Ils permettent de valider ou d'infirmer l'hypothèse selon laquelle les variables ne sont pas significativement corrélées.

Mesure de précision de l'échantillonnage de Kaiser-Meyer-Olkin : ce tableau donne pour chaque variable la valeur de la mesure KMO ainsi que le KMO global. L'indice KMO varie entre 0 et 1. Une valeur faible correspond au cas où il n'est pas possible d'extraire de facteurs synthétiques (ou variables latentes). Autrement dit, les individus ne permettent pas de faire ressortir le modèle que l'on pouvait imaginer préalablement (l'échantillon est « inadéquat »). Kaiser (1974) recommande de ne pas accepter une décomposition si le KMO est inférieur à 0.5. Si le KMO est entre 0.5 et 0.7 alors la qualité de l'échantillon est moyenne, elle est bonne pour un KMO entre 0.7 et 0.8, très bonne entre 0.8 et 0.9 et excellente au-delà.

Valeurs propres : les valeurs propres et le graphique (*scree plot*) correspondant sont affichés. Le nombre de valeurs propres est égal au nombre de valeurs propres non nulles.

Si les options de sorties correspondantes ont été activées, XLSTAT affiche ensuite les **coordonnées des variables** dans le nouvel espace, puis les corrélations entre les variables d'origine et les composantes dans le nouvel espace. Les **corrélations** sont égales aux coordonnées des variables dans le cas d'une ACP normée (sur matrice de corrélation).

Si des variables supplémentaires ont été sélectionnées les coordonnées et les corrélations correspondantes sont affichées en fin de tableau.

Contributions : les contributions sont une aide à l'interprétation. Les variables ayant le plus influencé la construction des axes sont celles dont les contributions sont les plus élevées.

Cosinus carrés : comme pour les autres méthodes factorielles, l'analyse des cosinus carrés permet d'éviter des erreurs d'interprétation dues à des effets de projection. Si les cosinus carrés associés aux axes utilisés sur un graphique sont faibles, on évitera d'interpréter la position de l'observation ou de la variable en question.

Les **coordonnées des observations** dans le nouvel espace sont ensuite affichées. Si des données supplémentaires ont été sélectionnées, elles sont affichées en fin de tableau.

Contributions : ce tableau fournit les contributions des observations à la construction des composantes principales.

Cosinus carrés : dans ce tableau sont affichés les cosinus carrés entre les vecteurs observations et les axes factoriels.

Dans le cas où une rotation a été demandée, les résultats de la rotation sont affichés, avec en premier la **matrice de rotation** appliquée aux coordonnées des variables. Suivent ensuite les pourcentages modifiés de variabilité associés à chacun des axes concernés par la rotation. Dans les tableaux suivants sont affichées les coordonnées, les contributions et les cosinus des variables et des observations après rotation.

Exemple

Un exemple d'utilisation de la cartographie interne des préférences est disponible sur le Centre d'aide XLSTAT à l'adresse

http://www.xlstat.com/demo-intprefmapf.htm

Bibliographie

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research*, **1**, 245-276.

Gabriel K.R. (1971). The biplot graphic display of matrices with application to principal component analysis. *Biometrika*, **58**, 453-467.

Gower J.C. and Hand D.J. (1996). Biplots. Chapman and Hall, London.

Jobson J.D. (1992). Applied multivariate data analysis. Volume II: Categorical and Multivariate Methods. Springer-Verlag, New York.

Jolliffe I.T. (2002). Principal Component Analysis, Second Edition. Springer, New York.

Kaiser H. F. (1974). An index of factorial simplicity. *Psychometrika*, **39**, 31-36.

Legendre P. and Legendre L. (1998). Numerical Ecology. Second English Edition. Elsevier, Amsterdam, 403-406.

Mauchly J. W. (1940). Significance test for sphericity of a normal n-variate distribution. The *Annals of Mathematical Statistics*. **11**, 204-209.

Morineau A. and Aluja-Banet T. (1998). Analyse en Composantes Principales. CISIA-CERESTA, Paris.

Rao C. R. (1964). The use and interpretation of principal components analysis in applied research. *Sankhya*, A **26**, 329-358.

Analyse de données de préférences

Utilisez cette fonction pour analyser des données de préférences rapidement et efficacement.

Cette fonction permet de :

- déterminer les produits les plus appréciés
- faire des comparaisons de produits
- réaliser des comparaisons entre sujets ou groupes de sujets

Dans cette section :

Description

Les données de préférences (dites aussi données hédoniques ou données de liking) font parties des plus collectées en analyse sensorielle. Elles consistent simplement à demander aux différents sujets/consommateurs de donner une note aux produits, avec généralement une échelle prédéfinie sur laquelle ces derniers doivent répondre.

Si le principe des données de préférences est très simple, l'analyse de ces données est assez riche. La première étape est une description des données de liking, avec leur distribution par produit, les différences entre les sessions, la visualisation des données... Une seconde étape, plus poussée, consiste à réaliser des tests de comparaisons entre les produits ainsi qu'à construire une cartographie interne des préférences. La dernière étape est basée sur l'étude des accords entre les sujets avec la comparaison de groupes de sujets ou encore la classification de ces derniers.

Structure des données

Il existe deux formats différents :

1. Toutes les données des sujets sont concaténées horizontalement (format horizontal).
2. Toutes les données sont concaténées verticalement (format vertical).

Pour la saisie des données, XLSTAT vous demande de sélectionner l'ensemble des données, et de donner le type de format. Dans le cas du format vertical, les produits et les sujets sont demandés. Il est à noter que si vous rentrez plusieurs colonnes en format vertical, ces dernières seront moyennées. De même, en cas de sessions, ces dernières seront étudiées puis moyennées afin de se ramener au format horizontal.

Valeurs manquantes

Lors d'un format vertical entré, XLSTAT ramènera automatiquement les données au format horizontal. Ce qui implique les remarques suivantes :

- Si plusieurs colonnes sont rentrées et que toutes les valeurs ne sont pas manquantes sur une ligne, alors la moyenne est réalisée sur les valeurs non manquantes et nous n'avons pas de valeur manquante au format horizontal.
- Si un sujet n'a pas vu un produit pour une session, alors la moyenne est réalisée sur les sessions existantes. Il n'y aura donc pas de valeur manquante au format horizontal non plus.
- Si un sujet n'a pas du tout vu un produit, que ce soit signalé par une valeur manquante dans les données ou simplement par l'absence de la combinaison Produit x Sujet, alors une valeur manquante sera présente dans le format horizontal (elle sera estimée si c'est l'option choisie).

Enfin, il est à noter que tous les résultats affichés avant la fin du pré-traitement des données (Description des sessions, visualisation des données, moyennes des produits et des sujets avant centrage/réduction) tiennent compte des valeurs manquantes, au contraire des résultats qui suivent le pré-traitement des données.

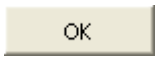
Classification des sujets

Il est possible de réaliser une classification des sujets. Cette dernière se fait grâce à une Classification Ascendante Hiérarchique basée sur la distance euclidienne et le critère de Ward. Dans le cas où un choix automatique du nombre de classes est demandé, l'indice d'Hartigan est utilisé.

Si vous voulez comparer les groupes obtenus par la classification des sujets, il suffit de sélectionner vos données de préférences au format horizontal (qui sont soit vos données initiales soit les données affichées par XLSTAT) et le vecteur de résultat des groupes dans le champ "Groupes de sujets".

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.

 : cliquez sur ce bouton pour effacer les sélections de données.

 : cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.

 : cliquez sur ce bouton pour changer la façon dont vous chargez les données dans XLSTAT. Si le bouton possède une icône de souris, XLSTAT vous permet de faire une sélection à la souris des données. Si le bouton possède une icône de liste, XLSTAT vous permet de sélectionner les données à partir d'une liste. Si le bouton possède une icône orange, des boutons complémentaires avec un point d'interrogation  sont affichés pour vous permettre d'importer les données à partir de fichiers.

Onglet **Général** :

Format : Cliquez sur horizontal ou vertical selon la façon dont vos données sont structurées.

Données de préférences : sélectionnez les données correspondant aux différents sujets. Si la première ligne de la sélection comprend des en-têtes, l'option « Libellés des variables » en format vertical ou « Libellés des sujets » en format horizontal doit être activée. Si vous êtes en format vertical et que vous sélectionnez plusieurs colonnes, ces dernières seront moyennées.

Si le format est **horizontal** :

Libellés des produits : activez cette option si vous voulez utiliser des libellés des produits pour l'affichage des résultats. Si l'option « Libellés des sujets » est activée, la première cellule de la sélection doit comprendre un en-tête. Si vous n'activez pas cette option, des libellés seront automatiquement créés.

Si le format est **vertical** :

Produits : sélectionnez les produits correspondants aux lignes des données de préférences. Si l'option « Libellés des variables » est activée, la première cellule de la sélection doit comprendre un en-tête.

Sujets : sélectionnez les sujets correspondants aux lignes des données de préférences. Si l'option « Libellés des variables » est activée, la première cellule de la sélection doit comprendre un en-tête.

Sessions : sélectionnez les sessions correspondants aux lignes des données de préférences. Si l'option « Libellés des variables » est activée, la première cellule de la sélection doit comprendre un en-tête.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des sujets (ou des variables) : activez cette option si la première ligne des données sélectionnées (Données de préférences, Libellés des produits, Sujets, Sessions, Groupes de sujets) contient un libellé. Si vous n'activez pas cette option, des libellés seront automatiquement créés.

Groupes de sujets : Sélectionnez les données correspondant aux groupes de sujets. Elles doivent comporter autant de valeurs que de sujets (nombre de colonnes dans les données).

Onglet **Options** :

Centrer les sujets : Activez cette option pour que les sujets soient centrés (moyenne de chaque sujet ramenée à 0).

Réduire les sujets : Activez cette option pour que les sujets soient réduits (variance de chaque sujet ramenée à 1).

Intervalle de confiance (%) : entrez l'étendue en pourcentage de l'intervalle de confiance à utiliser pour les différents tests, et pour le calcul des intervalles de confiance. Valeur par défaut : 95.

Classification des sujets : Activez cette option pour réaliser une classification des sujets (détails section "Classification des sujets"). Dans un second temps, déterminez si vous voulez que XLSTAT définisse **automatiquement** une troncature, et donc le nombre de classes à retenir, ou si vous voulez définir vous-même le **nombre de classes** à créer.

Onglet **Données manquantes** :

Des détails sont donnés section "Valeurs manquantes".

Ne pas accepter les valeurs manquantes : activez cette option pour que XLSTAT empêche la poursuite des calculs si des valeurs manquantes au format horizontal sont détectées

Estimer les données manquantes : activez cette option pour estimer les données manquantes avant le début des calculs.

- **Moyenne du produit** : activez cette option pour estimer les données manquantes en utilisant la moyenne du produit correspondant.
- **Moyenne du sujet** : activez cette option pour estimer les données manquantes en utilisant la moyenne du sujet correspondant.

Onglet **Sorties** :

Différences entre les sessions : activez cette option pour afficher les différences entre les sessions. Si vous avez plus de deux sessions, ce seront les écarts-types entre les sessions qui seront affichés.

Données au format horizontal : activez cette option pour afficher les données au format horizontal.

Moyennes des produits : activez cette option pour afficher le tableau des moyennes des produits.

Moyennes des sujets : activez cette option pour afficher le tableau des moyennes des sujets.

Tests sur les moyennes des produits : activez cette option pour afficher les résultats des tests sur les moyennes des produits (ANOVA et tests de comparaisons multiples).

Cartographie interne des préférences : activez cette option pour afficher les tableaux des résultats provenant de la cartographie interne des préférences.

Différences entre les groupes : activez cette option pour afficher les résultats des tests sur les moyennes des groupes produit par produit (ANOVA et tests de comparaisons multiples).

Interprétation : activez cette option pour que XLSTAT calcule une interprétation automatique des résultats des ANOVAs.

Composition des classes : activez cette option pour afficher la composition des classes obtenue après troncature du dendrogramme.

Onglet **Graphiques** :

Différences entre les sessions : activez cette option pour afficher les graphiques des différences entre les sessions. Si vous avez plus de deux sessions, ce seront les écarts-types entre les sessions qui seront affichés.

Box plots : activez cette option pour afficher le box plot de chacun des produits.

Moyennes des produits : activez cette option pour afficher le graphique des moyennes des produits.

Moyennes des sujets : activez cette option pour afficher le graphique des moyennes des sujets.

Visualisation des données : activez cette option pour afficher le graphique permettant de visualiser les données des différents sujets.

Graphiques des moyennes : activez cette option pour afficher les graphiques permettant de visualiser les résultats des tests de comparaisons multiples entre les produits.

Cartographie interne des préférences : activez cette option pour afficher les graphiques provenant de la cartographie interne des préférences.

Différences entre les groupes : activez cette option pour afficher les graphiques permettant de visualiser les résultats des tests de comparaisons multiples entre les groupes.

Dendrogramme : activez cette option pour afficher le dendrogramme.

- **Horizontal** : choisissez cette option pour afficher un dendrogramme horizontal.
- **Vertical** : choisissez cette option pour afficher un dendrogramme vertical.

- **Tronqué** : activez cette option pour afficher le dendrogramme tronqué (le dendrogramme commence au niveau de la troncature).
- **Etiquettes** : activez cette option pour afficher les libellés des sujets (dendrogramme complet) ou des classes (dendrogramme tronqué) sur le dendrogramme.
- **Couleurs** : activez cette option pour utiliser des couleurs pour représenter les différents groupes sur le dendrogramme complet.

Résultats

Différences entre les sessions : le tableau des différences entre les sessions (écarts-types si vous avez plus de deux sessions) est affiché. Il permet de voir les éventuelles erreurs des sujets ou de saisie (notamment si une valeur est élevée).

Moyenne des différences entre sessions pour chaque produit : le tableau des moyennes des différences entre les sessions par produit est affiché suivi du graphique associé. Ces résultats permettent de déterminer si certains produits ont donné lieu à des écarts entre les sessions.

Moyenne des différences entre sessions pour chaque sujet : le tableau des moyennes des différences entre les sessions par sujet est affiché suivi du graphique associé. Ces résultats permettent de déterminer si certains sujets ont donné lieu à des écarts entre les sessions.

Données au format horizontal : les données au format horizontal sont affichées. Ces dernières sont sans données manquantes (elles sont estimées par l'option choisie). Ces données vous permettent de choisir vous-même certaines options en les rentrant dans une cartographie interne des préférences, une classification ascendante hiérarchique...

Si vous avez sélectionné des groupes, les résultats suivants seront affichés groupe par groupe. De plus, si vous avez sélectionné l'option "Centrer les sujets", certains résultats seront donnés avant et après centrage des sujets.

Moyennes des produits : le tableau des moyennes des produits ainsi que le diagramme en bâtons associé sont affichés. Ce résultat permet de déterminer à quel point les produits sont appréciés.

Box plots des données de préférences par produit : les box plots des données de préférences pour chaque produit sont affichés. Ces derniers permettent de visualiser la dispersion des données de préférences au sein d'un produit et de comparer les dispersions entre les produits.

Visualisation des données : un graphique permettant de visualiser directement les données des différents sujets est affiché. Vous pouvez choisir le sujet à mettre en lumière afin de vérifier ses données ou de le comparer aux autres.

ANOVA : ce tableau permet d'évaluer le pouvoir explicatif du facteur produit. Le pouvoir explicatif est évalué en comparant l'ajustement (au sens des moindres carrés) du modèle final avec l'ajustement du modèle rudimentaire composé d'une constante égale à la moyenne de la variable dépendante (Données de préférences). En d'autres termes, si la p-value est significative, nous rejetons l'hypothèse stipulant que toutes les moyennes des produits sont égales.

Graphiques des moyennes : ces graphiques permettent de comparer visuellement les moyennes des produits avec les intervalles de confiance associés.

Produit/Tukey (HSD) : les résultats des tests de comparaisons multiples des moyennes des produits sont affichés, afin de déterminer les produits différents les uns des autres ainsi que ceux similaires. Les groupes des produits sont ensuite donnés.

Cartographie interne des préférences : les résultats de la cartographie interne des préférences sont affichés. Ils démarrent par les valeurs propres des facteurs ainsi que les pourcentages d'inertie que chacun représente, avant d'afficher les coordonnées des sujets et les coordonnées des produits. Toutes ces coordonnées sont également affichées dans des graphiques. Remarque : si un sujet n'a pas une qualité de représentation supérieure à 50% (somme des cosinus carrés du sujet sur les axes > 0.5), alors il n'est pas affiché.

Différences pour chaque produit : les résultats de l'ANOVA, des tests de comparaisons multiples entre classes et les graphiques associés sont affichés pour chacun des produits.

Classification des sujets : les résultats de la classification des sujets sont affichés. Ils se composent tout d'abord du dendrogramme obtenu, éventuellement du dendrogramme tronqué si celui-ci a été demandé, et des classes des sujets construites par la coupure du dendrogramme.

Exemple

Un exemple d'utilisation d'analyse de données de préférences est disponible sur le Centre d'aide XLSTAT à l'adresse suivante :

http://www.xlstat.com/demo-likf.htm

Bibliographie

Hsu J.C. (1996). Multiple Comparisons: Theory and Methods. CRC Press, Boca Raton.

Jolliffe I.T. (2002). Principal Component Analysis, Second Edition. Springer, New York.

Ward J.H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, **58**, 238-244.

Analyse de Panel

Utilisez cet outil pour tester si votre panel de consommateurs ou d'experts permet de mettre en évidence des différences entre les produits évalués et si oui, dans quelle mesure. Vérifiez également si les notes données par le panel peuvent être considérées comme fiables ou non.

Dans cette section :

Description

Cet outil permet d'enchaîner différentes analyses proposées par XLSTAT, afin d'évaluer la capacité d'un panel de J consommateurs, experts, sujets, ou assesseurs (le terme sujet est utilisé dans l'interface de XLSTAT), à différencier P produits suivant K descripteurs (des variables au sens statistique) et à contrôler si les notes sont fiables (si des mesures répétées sont disponibles) et si l'on peut identifier des groupes homogènes parmi les sujets.

La **première étape** consiste en une série d'ANOVA dont le but est de vérifier pour chaque descripteur s'il permet de mettre en évidence un effet produit ou non. Pour chaque descripteur, le tableau des ANOVA de Type III pour le modèle choisi est affiché. Un tableau de synthèse permet de comparer les p-values associées au facteur produit pour les différents descripteurs. Si vous avez demandé un filtrage, la suite de l'analyse ne sera conduite que pour les descripteurs permettant de différencier les produits. Différents modèles d'ANOVA sont possibles en fonction de la présence ou non de sessions, de la volonté de prendre en compte les interactions dans le modèle et de la volonté de considérer les effets sujets et session comme fixes ou aléatoires.

Table CAP

La table CAP comporte deux parties. La partie gauche est un résumé des descripteurs. Ces derniers sont triés en fonction de leur discrimination des produits. Si la p-value est inférieure à 0.1, la couleur sera jaune. Si elle est inférieure à 0.05, la couleur sera verte. Autrement, la couleur sera rouge. C'est exactement le contraire pour l'interaction $produit \times sujet$ puisque ce n'est pas une chose positive d'avoir une interaction significative. La moyenne de l'attribut et la racine carrée de l'erreur terminent cette partie gauche. La partie droite de du tableau concerne les sujets. Attention, si un filtrage a été effectué, cette partie sera affichée sous la précédente. Les sujets sont triés en fonction de leur moyenne des rangs des effets produits individuels sur l'ensemble des descripteurs. Pour un descripteur donné, si un sujet ne discrimine pas les produits, il aura alors un « = ». S'il les discrimine, soit il est en accord avec le panel (test sur sa contribution à l'interaction sujet*produit) et aura un « + », sinon il aura un « -

». Enfin, si le sujet a un effet session pour le descripteur en question (effet d'humeur), ou si il est significativement moins fiable que les autres sujets d'une session à l'autre, il est considéré non répétable et aura alors un « ! » ajouté.

La **seconde étape** consiste en une analyse graphique. Pour chacun des k descripteurs retenus sont affichés des box plots et des strip plots. On peut ainsi visualiser comment, pour chaque descripteur, les différents sujets utilisent l'échelle de notation pour évaluer les différents produits.

La **troisième étape** consiste en une restructuration du tableau de données, afin d'avoir un tableau contenant une ligne par produit et une colonne par couple de sujet et descripteur (s'il y a des sessions, le tableau contient alors les moyennes), puis en une ACP (normée) sur ce tableau. Le nombre de produits P étant en général inférieur au produit $k \times J$, on a au plus P axes factoriels. On affiche ensuite autant de cercles de corrélation qu'il y a de descripteur, en faisant ressortir pour chaque descripteur les couples (sujet, descripteur) où il est impliqué. On peut ainsi vérifier en une étape dans quelle mesure les sujets sont d'accord ou non pour les k descripteurs, une fois l'effet de position et d'échelle supprimé (car l'ACP est normée), et dans quelle mesure les descripteurs sont liés ou non. Afin d'étudier plus précisément la relation entre les descripteurs, une AFM (Analyse Factorielle Multiple) est réalisée.

Au cours de la **quatrième étape** est réalisée pour chaque sujet, une ANOVA pour chacun des k descripteurs afin de vérifier s'il y a un effet produit. Cela permet d'évaluer pour chaque sujet sa capacité à distinguer les produits au travers des critères/descripteurs utilisés. Voir la section description de la méthode ANOVA pour plus de détails. Un tableau de synthèse permet ensuite de compter pour chaque sujet le nombre de descripteurs pour lesquels il a pu faire la différence entre les produits et le pourcentage correspondant est affiché. Ce pourcentage est une mesure simple du pouvoir discriminant des sujets.

Pour la **cinquième étape**, un tableau global présente dans un premier temps les notes (moyennées sur les sessions éventuelles) pour chaque sujet en ligne et chaque couple (produit, descripteur) en colonne. Il est suivi d'une série de P tableaux et graphiques permettant, pour chaque produit pris séparément, de comparer les sujets (moyennés sur les sessions éventuelles) pour l'ensemble des descripteurs. Ces graphiques permettent d'identifier des tendances fortes et d'éventuelles notations atypiques pour certains sujets.

La **sixième étape** permet de repérer les sujets atypiques au travers de la mesure pour chaque produit d'une distance euclidienne de chaque sujet à une moyenne calculée pour sur l'ensemble des sujets dans l'espace des descripteurs. Un tableau affichant ces distances ainsi que pour chaque produit le minimum et le maximum, permet d'identifier les sujets proches ou éloignés du consensus. Un graphique permet ensuite de visualiser ces distances.

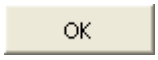
Si une variable session a été sélectionnée, la **septième étape** permet de vérifier si pour certains sujets il y a un effet ordre de session, au travers d'un test de Friedman (ou de Wilcoxon signé s'il n'y a que deux sessions) calculé sur l'ensemble des produits, descripteur par descripteur. On calcule ensuite pour chaque sujet et chaque descripteur, quelle est l'amplitude maximale observée entre les sessions. Le produit correspondant à l'amplitude maximale est indiqué sur le triangle rouge. Ce tableau permet de repérer d'éventuelles anomalies dans les notes données par certains sujets et éventuellement de supprimer certaines observations pour des analyses futures.

Si pour chaque triplet (sujet, produit, descripteur) il existe au moins une note, la **huitième étape** consiste en une classification des sujets. La classification est d'abord réalisée sur les données non centrées-réduites, puis sur les données centrées réduites, afin de supprimer les éventuels effets d'échelle et de position.

Enfin un tableau préformaté pour utiliser la méthode STATIS est présent. Cette méthode vous permettra d'avoir des indices d'accords entre les sujets et plus généralement d'un sujet avec le point de vue global du panel. De plus, une carte des produits sera réalisée. Voir la section description de la méthode STATIS pour plus de détails.

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.


: cliquez sur ce bouton pour changer la façon dont vous chargez les données dans XLSTAT. Si le bouton possède une icône de souris, XLSTAT vous permet de faire une sélection à la souris des données. Si le bouton possède une icône de liste, XLSTAT vous permet de sélectionner les données à partir d'une liste. Si le bouton possède une icône orange, des boutons complémentaires avec un point d'interrogation  sont affichés pour vous permettre d'importer les données à partir de fichiers.

Onglet **Général**:

Y / Descripteurs : sélectionnez les données associées aux notes données aux descripteurs. Le tableau doit contenir les notes attribuées par les sujets aux caractéristiques étudiées. Si des en-têtes de colonnes ont été sélectionnés, veuillez vérifier que l'option « Libellés des variables » est activée.

Produits : sélectionnez les données qui correspondent aux produits testés. Cette sélection ne peut contenir qu'une seule colonne. Si des en-têtes de colonnes ont été sélectionnés, veuillez vérifier que l'option « Libellés des variables » est activée.

Sujets : sélectionnez les données qui correspondent aux sujets. Cette sélection ne peut contenir qu'une seule colonne. Si des en-têtes de colonnes ont été sélectionnés, veuillez vérifier que l'option « Libellés des variables » est activée.

Sessions : activez cette option si plusieurs sessions de test ont eu lieu. Si tel est le cas, sélectionnez les données qui correspondent à la session. Cette sélection ne peut contenir qu'une seule colonne. Si des en-têtes de colonnes ont été sélectionnés, veuillez vérifier que l'option « Libellés des variables » est activée.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des variables : activez cette option si la première ligne des données sélectionnées (Descripteurs, Produits, Sujets, Sessions, Libellés des observations, Poids des observations) contient un libellé.

Libellés des observations : activez cette option si vous voulez utiliser des libellés pour les observations pour l'affichage des résultats. Si l'option « Libellés des variables » est activée, la première cellule de la sélection doit comprendre un en-tête. Si vous n'activez pas cette option, des libellés seront automatiquement créés (Obs1, Obs2, …).

Poids des observations : activez cette option si vous voulez pondérer les observations. Si vous n'activez pas cette option, les poids seront tous considérés comme valant 1. Les poids doivent être impérativement supérieurs ou égaux à 0. Si un en- tête de colonne a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

Onglet **Options**:

Modèle : sélectionnez le modèle d'analyse de la variance (ANOVA) qui sera utilisé notamment pour identifier les descripteurs non discriminants. Si l'option session n'est pas sélectionnée, les deux modèles proposés sont :

- $Y = Produit + Sujet$
- $Y = Produit + Sujet + Produit \times Sujet$

Si l'option session est sélectionnée, les trois modèles proposés sont :

- $Y = Produit + Sujet + Session$
- $Y = Produit + Sujet + Session + Produit \times Sujet$
- $Y = Produit + Sujet + Session + Produit \times Sujet + Produit \times Session + Session \times Sujet$

Effets aléatoires (Sujet / Session) : activez cette option si vous voulez considérer que les effets Sujet et Session et les éventuelles interactions les impliquant soient considérés comme des effets aléatoires. Si cette option n'est pas activée, tous les effets sont considérés comme fixes.

Niveau de signification (%) : entrez le niveau de signification utilisé pour déterminer en-deçà de quelle p-value, les différents tests amènent à rejeter l'hypothèse nulle associée.

Filtrer les descripteurs : activez cette option pour éliminer des analyses les descripteurs pour lesquels il n'y a pas d'effet produit. Entrez alors la p-value seuil au-delà de laquelle l'effet produit est considéré comme non significatif. Pour que cette opération soit réalisée,il ne faut pas que vous ayez décoché "Résumé des ANOVA" de la section Sorties.

Onglet **Données manquantes** :

Supprimer les observations : activez cette option pour supprimer les observations comportant des données manquantes.

- **Vérifier chaque Y séparément** : activez cette option pour supprimer les observations comportant des données manquantes en prenant chaque descripteur séparément (on pourra avoir des nombres d'observations différents d'un descripteur à un autre).
- **Pour tous les Y** : activez cette option pour supprimer toutes les observations comportant au moins une valeur manquante.

Estimer les données manquantes : activez cette option pour estimer les données manquantes avant le début des calculs.

- **Moyenne ou mode** : activez cette option pour estimer les données manquantes en utilisant la moyenne (variables quantitatives) ou le mode (variables qualitatives) pour les variables correspondantes.
- **Plus proche voisin** : activez cette option pour estimer les données manquantes d'une observation en recherchant le plus proche voisin de l'observation.

Onglet **Sorties**:

Statistiques descriptives : activez cette option pour afficher les statistiques descriptives pour les variables sélectionnées.

Résumés des ANOVA : activez cette option pour afficher les tableaux de synthèse des différentes ANOVA effectuées.

Capacité des sujets à différencier les produits : activez cette option pour afficher les tableaux et les graphiques permettant d'évaluer la capacité des sujets à différencier les produits.

Moyennes des sujets par (produit, descripteur) : activez cette option pour afficher le tableau des moyennes par couple (produit, descripteur) et pour chaque produit, le tableaux des moyennes par sujet et par descripteur.

Distances au consensus : activez cette option pour afficher le tableau des distances au consensus.

Analyse des sessions : activez cette option pour évaluer la fiabilité des sujets sur la base des sessions.

Tableau pour STATIS : activez cette option pour afficher le tableau formaté pour effectuer une analyse STATIS.

Onglet **Graphiques** :

Box plots : activez cette option pour afficher les box plots permettant pour chacun des descripteurs de comparer les différents sujets.

Strip plots : activez cette option pour afficher les strip plots permettant pour chacun des descripteurs de comparer les différents sujets.

Graphiques d'ACP : activez cette option pour afficher les graphiques issus des ACP.

Graphiques en lignes par produit : activez cette option pour afficher les graphiques en ligne permettant pour chaque produit de comparer les sujets sur l'ensemble des descripteurs.

Graphique en lignes des distances au consensus : activez cette option pour afficher le graphique en ligne représentant pour chaque produit la distance au consensus de chaque sujet.

Dendrogramme : activez cette option pour afficher les dendrogrammes issus de la classification des sujets.

Résultats

Statistiques descriptives : le tableau des statistiques descriptives présente des statistiques simples pour toutes les variables sélectionnées. Le nombre de valeurs manquantes, le nombre de valeurs non manquantes, la moyenne, l'écart standard sont affichés pour les variables quantitatives. Pour les variables qualitatives, les catégories avec leur fréquence respectives et pourcentages sont affichés.

Première étape : Les premiers tableaux affichés correspondent aux ANOVA de Type III réalisées pour chaque descripteur afin de déterminer s'il y a ou non un effet produit.

Si l'option de filtrage a été activée, la suite de l'analyse ne concerne que les descripteurs permettant de différencier les produits.

Viens ensuite la table CAP (Control of Assessor Performances). Voir la section [description] pour plus de détails.

Deuxième étape : Pour chacun des descripteurs retenus sont affichés des box plots et des strip plots. On peut ainsi visualiser comment, pour chaque descripteur, les différents sujets utilisent l'échelle de notation pour évaluer les différents produits.

La **troisième étape** consiste en une restructuration du tableau de données, afin d'avoir un tableau contenant une ligne par produit et une colonne par couple de sujet et descripteur (s'il y a des sessions, le tableau contient alors les moyennes), puis en une ACP (normée) sur ce tableau. On affiche ensuite autant de cercles de corrélation qu'il y a de descripteur, en faisant ressortir pour chaque descripteur les couples (sujet, descripteur) où il est impliqué. Afin d'étudier plus précisément la relation entre les descripteurs, une AFM est réalisée.

Au cours de la **quatrième étape** est réalisée pour chaque sujet, une ANOVA pour chacun des k descripteurs afin de vérifier s'il y a un effet produit. Un tableau de synthèse permet ensuite de compter pour chaque sujet le nombre de descripteurs pour lesquels il a pu faire la différence entre les produits et le pourcentage correspondant est affiché. Ce pourcentage est une mesure simple du pouvoir discriminant des sujets.

Pour la **cinquième étape**, un tableau global présente dans un premier temps les notes (moyennées sur les sessions éventuelles) pour chaque sujet en ligne et chaque couple (produit, descripteur) en colonne. Il est suivi d'une série de P tableaux et graphiques permettant, pour chaque produit pris séparément, de comparer les sujets (moyennés sur les sessions éventuelles) pour l'ensemble des descripteurs. Ces graphiques permettent d'identifier des tendances fortes et d'éventuelles notations atypiques pour certains sujets.

La **sixième étape** permet de repérer les sujets atypiques au travers de la mesure pour chaque produit d'une distance euclidienne de chaque sujet à une moyenne calculée pour sur l'ensemble des sujets dans l'espace des descripteurs. Un tableau affichant ces distances ainsi que pour chaque produit le minimum et le maximum, permet d'identifier les sujets proches ou éloignés du consensus. Un graphique permet ensuite de visualiser ces distances.

Si une variable session a été sélectionnée, la **septième étape** permet devérifier si pour certains sujets il y a un effet ordre de session, au travers d'un test de Friedman (ou de Wilcoxon signé s'il n'y a que deux sessions) calculé sur l'ensemble des produits, descripteur par descripteur. On calcule ensuite pour chaque sujet et chaque descripteur, quelle est l'amplitude maximale observée entre les sessions. Le produit correspondant à l'amplitude maximale est indiqué sur le triangle rouge.

Si pour chaque triplet (sujet, produit, descripteur) il existe au moins une note, la **huitième étape** consiste en une classification des sujets. La classification est d'abord réalisée sur les données non centrées-réduites, puis sur les données centrées réduites, afin de supprimer les éventuels effets d'échelle et de position.

Enfin un tableau préformaté pour utiliser la méthode STATIS est présent.

Exemple

Un exemple d'utilisation de l'Analyse de Panel est disponible sur le Centre d'aide XLSTAT à l'adresse

http://www.xlstat.com/demo-panelf.htm

Bibliographie

Conover W.J. (1999). Practical Nonparametric Statistics, 3rd edition, Wiley.

Escofier B. and Pagès J. (1998). Analyses Factorielles Simples et Multiples : Objectifs, Méthodes et Interprétation. Dunod, Paris.

Næs T., Brockhoff P. and Tomic O. (2010). Statistics for Sensory and Consumer Science. Wiley, Southern Gate.

Caractérisation de produits

Utilisez cet outil pour identifier quels sont les descripteurs qui discriminent le mieux les produits et quelles sont les caractéristiques importantes de ces mêmes produits dans le cadre de l'analyse sensorielle.

Dans cette section :

Description

Cet outil, développé suivant les recommandations de Jérôme Pagès et Sébastien Lê du Laboratoire de Mathématiques Appliquées de l'Agrocampus de Rennes, a pour but de permettre aux utilisateurs de XLSTAT de disposer d'un moyen rapide et rigoureux pour identifier quels sont les descripteurs discriminants d'une série de produits évalués lors d'une étude sensorielle et quelles sont les caractéristiques importantes des différents produits.

Les calculs réalisés s'appuient principalement sur l'ANOVA (analyse de variance) pour la modélisation. Pour plus de détails techniques, voir le chapitre sur l'analyse de la variance de l'aide d'XLSTAT.

Le tableau de données utilisé pour l'analyse doit être constitué de lignes donnant pour un produit donné et éventuellement une session donnée, la note attribuée par un sujet donné pour l'ensemble des descripteurs (ou caractéristiques) étudiée. On aura donc trois colonnes indiquant l'identifiant du sujet, l'identifiant du produit, éventuellement de la session, et autant de colonnes que de descripteurs (ou caractéristiques) évalués par les sujets.

Pour chacune des caractéristiques, une ANOVA est réalisée afin de déterminer si les notes attribuées par les sujets sont significativement différentes ou non. Le modèle le plus simple est

Note descripteur = effet produit + effet sujet

Si des répétitions sont disponibles, i.e. si chaque sujet a évalué au moins deux fois chaque produit, on pourra ajouter le facteur session dans le modèle, donnant alors le modèle

Note descripteur = effet produit + effet sujet + effet session

On peut aussi ajouter les interactions. Cet ajout permet alors de tester si certaines combinaisons (produit, sujet) ont tendance à donner des notes plus fortes ou au contraire moins fortes. On a alors le modèle :

Note descripteur = effet produit + effet sujet + effet produit * effet sujet

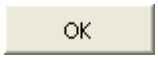
On considère que l'effet sujet est aléatoire. Cela signifie que l'on considère que chaque sujet a sa propre tendance à utiliser plus ou moins largement l'échelle de notation. Cette tendance peut varier d'un descripteur à l'autre, les ANOVA réalisées pour les différents descripteurs étant indépendantes.

La caractérisation de produits permet de caractériser rapidement des produits en fonction des préférences des sujets.

Les modèles sont eux-mêmes construits à partir de données objectives (par exemple des descripteurs physico-chimiques, ou des notes fournies par des experts sur des critères bien déterminés) ce qui permet d'interpréter la position des produits en fonction de critères objectifs.

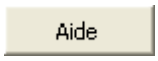
Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.



: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.

Onglet **Général**:

Y / Descripteurs : sélectionnez les données associées aux notes données aux descripteurs. Le tableau doit contenir les notes attribuées par les sujets aux caractéristiques étudiées. Si des en-têtes de colonnes ont été sélectionnés, veuillez vérifier que l'option « Libellés des variables » est activée.

Produits : sélectionnez les données qui correspondent aux produits testés. Cette sélection ne peut contenir qu'une seule colonne. Si des en-têtes de colonnes ont été sélectionnés, veuillez vérifier que l'option « Libellés des variables » est activée.

Sujets : sélectionnez les données qui correspondent aux sujets. Cette sélection ne peut contenir qu'une seule colonne. Si des en-têtes de colonnes ont été sélectionnés, veuillez vérifier que l'option « Libellés des variables » est activée.

Sessions : activez cette option si plusieurs sessions de test ont eu lieu. Si tel est le cas, sélectionnez les données qui correspondent à la session. Cette sélection ne peut contenir qu'une seule colonne. Si des en-têtes de colonnes ont été sélectionnés, veuillez vérifier que l'option « Libellés des variables » est activée.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des variables : activez cette option si la première ligne des données sélectionnées (variables dépendantes et explicatives, libellés des observations) contient un libellé.

Libellés des observations : activez cette option si vous voulez utiliser des libellés pour les observations pour l'affichage des résultats. Si l'option « Libellés des variables » est activée, la première cellule de la sélection doit comprendre un en-tête. Si vous n'activez pas cette option, des libellés seront automatiquement créés (Obs1, Obs2, …).

Poids : activez cette option si vous voulez pondérer les observations. Si vous n'activez pas cette option, les poids seront tous considérés comme valant 1. Les poids doivent être impérativement supérieurs ou égaux à 0. Si un en- tête de colonne a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

Onglet **Options**:

Modèle : sélectionnez le modèle d'analyse de la variance (ANOVA) qui sera utilisé notamment pour identifier les descripteurs non discriminants. Si l'option session n'est pas sélectionnée, les deux modèles proposés sont **Y = Produit + Sujet** et **Y= Produit + Sujet + Produit * Sujet**. Si l'option *Sessions* est sélectionnée, les trois modèles proposés sont **Y = Produit + Sujet + Session**, **Y = Produit + Sujet + Session + Produit * Sujet**, **Y = Produit + Sujet + Session + Produit * Sujet + Produit * Session + Session * Sujet**.

Trier le tableau des moyennes ajustées : activez cette option si vous désirez que les moyennes ajustées soient organisées de manière à ce que les produits et les descripteurs similaires soient le plus proche possible. Une analyse en composantes principales est utilisée afin de trouver le meilleur positionnement.

Niveau de signification (%) : entrez le niveau de signification pour les intervalles de confiance.

Onglet **Données manquantes** :

Supprimer les observations : activez cette option pour supprimer les observations comportant des données manquantes.

- **Vérifier chaque Y séparément** : activez cette option pour supprimer les observations comportant des données manquantes en prenant chaque descripteur séparément (on pourra avoir des nombres d'observations différent d'un descripteur à un autre).
- **Pour tous les Y** : activez cette option pour supprimer toutes les observations comportant au moins une valeur manquante.

Estimer les données manquantes : activez cette option pour estimer les données manquantes avant le début des calculs.

- **Moyenne ou mode** : activez cette option pour estimer les données manquantes en utilisant la moyenne (variables quantitatives) ou le mode (variables qualitatives) pour les variables correspondantes.
- **Plus proche voisin** : activez cette option pour estimer les données manquantes d'une observation en recherchant le plus proche voisin de l'observation.

Onglet **Graphiques**:

Profils sensoriels : activez cette option pour afficher le graphique des profils sensoriels.

- **Biplot** : activez cette option pour afficher simultanément les produits et les variables Y (descripteurs).
- **Filtrer les descripteurs non discriminants** : activez cette option pour ne pas prendre en compte les descripteurs qui sont apparus comme non discriminants lors des ANOVA. Vous pouvez saisir le **seuil** de probabilité au-delà duquel les descripteurs sont supprimés.

Résultats

Statistiques descriptives : les tableaux de statistiques descriptives présentent pour toutes les variables sélectionnées des statistiques simples. Pour les descripteurs (qui sont des variables quantitatives), sont affichés le nombre d'observations, le nombre de données manquantes, le nombre de données non manquantes, la moyenne, et l'écart-type (non biaisé). Pour les variables qualitatives sont affichés le nom des différentes modalités ainsi que leur fréquence respective.

Pouvoir discriminant par descripteur : dans ce tableau sont affichées les descripteurs ordonnés de celui qui a le plus fort pouvoir discriminant sur les produits à celui qui a le plus faible. Les valeurs du V-test ainsi que la p-value sont aussi affichées. Un graphique des p-values obtenues est affiché ensuite.

Coefficients du modèle : dans ce tableau sont affichés, pour chaque descripteur et pour chaque produit, les coefficients du modèle sélectionné. Pour chaque combinaison descripteur-produit, le coefficient, la moyenne estimée, la p-value ainsi qu'un intervalle de confiance sur le

coefficient sont affichés. Pour chaque produit, un graphique des coefficients associés aux différents descripteurs est affiché.

Moyennes ajustées par produit : ce tableau correspond aux moyennes ajustées calculées à partir du modèle pour chaque combinaison descripteur- produit. Les couleurs correspondent, pour le bleu, à un effet significativement positif du descripteur sur le produit et, pour le rouge, à un effet significativement négatif du descripteur sur le produit.

Graphique avec des ellipses de confiance pour les profils sensoriels obtenus par ACP : ce biplot, créé suivant la méthode décrite par Husson et al (2005) permet de visualiser sur un même graphique les descripteurs (après éventuel filtrage), ainsi que les produits avec une ellipse de confiance dont l'orientation et la surface dépend des notes données par les différents sujets. Ces ellipses sont calculées grâce à une méthode de rééchantillonnage. Les coordonnées des produits et les cosinus correspondants sont affichés afin d'éviter des erreurs d'interprétation.

Exemple

Un exemple de caractérisation de produit est disponible sur le Centre d'aide XLSTAT :

http://www.xlstat.com/demo-decatf.htm

Bibliographie

Husson F., Lê S. and Pagès J. (2009). SensoMineR dans Evaluation sensorielle - Manuel méthodologique. Lavoisier, SSHA, 3ème édition.

Husson F., Lê S. and Pagès J. (2005). Confidence ellipse for the sensory profiles obtained by principal component analysis. *Food Quality and Preference*, **16**, 245-250.

Lê S. and Husson F. (2008). SensoMineR: a package for sensory data analysis. *Journal of Sensory Studies*. **23(1)**. 14-25.

Lea P., Naes, T. and Rodbotten M. (1997). Analysis of Variance for Sensory Data. John Wiley, New York.

Naes T. and Risvik E. (1996). Multivariate Analysis of Data in Sensory Science. Elsevier Science, Amsterdam.

Sahai H. and Ageel M.I. (2000). The Analysis of Variance. Birkhaüser, Boston.

Penalty analysis

Utilisez cet outil pour analyser les résultats d'une enquête portant sur échelles de type JAR (*Just About Right*), pour lesquelles le niveau intermédiaire 3 correspond à la préférence du consommateur.

Dans cette section :

Description

La *penalty analysis* (analyse des pénalités) est une méthode utilisée en analyse sensorielle pour identifier des axes d'améliorations possibles pour des produits, suite à des enquêtes auprès de consommateurs ou d'experts.

Les données utilisées sont de deux types :

- des données de préférence correspondant à des indices de satisfaction globaux sur un produit (par exemple, une note d'appréciation globale de 1 à 10 pour un chocolat), ou sur une caractéristique d'un produit (le confort d'une voiture noté de 1 à 10) ;
- des données sur une échelle JAR (Just About Right) sur 5, 7 ou 9 niveaux. Dans le cas de 5 niveaux, ces données correspondent à des notes de 1 à 5 pour une ou plusieurs caractéristiques des produits étudiés où 1 correspond à « Pas du tout assez », 2 à « Pas assez », 3 à « JAR » (*Just About Right*) un idéal pour le consommateur, 4 à « Trop » et 5 à « Beaucoup trop ». Par exemple, pour un chocolat, on pourra noter son amertume, et pour le confort d'une voiture, le volume sonore du moteur.

La méthode consiste à identifier, en utilisant des ANOVA pour chacune des caractéristiques étudiées sur l'échelle JAR, si à une différence de notation JAR est associée une différence significative au niveau des données globales de préférence. Par exemple, le fait qu'un chocolat soit trop amer, est-il responsable d'un abaissement significatif de la note globale donnée à un chocolat ou non ?

Le terme de pénalité vient donc de ce que l'on recherche les caractéristiques susceptibles de pénaliser la satisfaction des consommateurs pour un produit donné. La pénalité est la différence de la moyenne des données de préférence pour la catégorie JAR, avec la moyenne des données pour les autres catégories.

L'analyse de pénalités se subdivise en trois phases :

1. On regroupe les données 1 et 2 d'une part et 4 et 5 d'autre part (dans le cas de l'échelle 1 à 5), ce qui permet d'une d'obtenir une échelle sur trois niveaux, « Pas assez », « JAR » et « Trop ».

2. On calcule puis on compare les moyennes des trois groupes pour les données de préférence pour identifier d'éventuelles différences significatives.

3. On calcule la pénalité puis on teste si elle est significativement différente de 0.

Tableau des pénalités : calculer la MCE et la différence standardisée

La MCE (en anglais : *Mean Squared Error (MSE)*) est un indicateur de la précision d'un modèle ou d'un système en comparant les valeurs prédites aux valeurs réelles. Le calcul de la MCE, qui est utilisé dans le calcul de certaines valeurs du tableau des pénalités (notamment pour calculer la différence standardisée), est légèrement différent du calcul de la MCE classique.

Calcul de la MCE pour le niveau JAR

$$MCE_{JAR} = \frac{\sum_{i=1}^{n}(y_i - MoyenneLikingJAR)^2 1_{Score=JAR} + (y_i - MoyenneLikingNonJA}{n-2}$$

où : - y_i correspond à la valeur de préférence (liking), - $MoyenneLikingJAR$ correspond à la moyenne des données de préférence ayant une note JAR, - $MoyenneLikingNonJAR$ correspond à la moyenne des données de préférence n'ayant pas une note JAR.

Calcul de la différence standardisée pour le niveau JAR

$$DifférenceStandardisée_{JAR} = \frac{Penalité}{\sqrt{(MCE(\frac{1}{n_{JAR}} + \frac{1}{n_{nonJAR}}))}}.$$

Calcul de la MCE pour le niveau Trop et Pas assez

$$MCE_{Trop/PasAssez} = \frac{\sum_{i=1}^{n}(y_i - MoyenneLikingJAR)^2 1_{Score=JAR} + (y_i - MoyenneLiking}{n -}$$

où : - y_i correspond à la valeur de préférence (liking), - $MoyenneLikingJAR$ correspond à la moyenne des données de préférence ayant une note JAR, - $MoyenneLikingTrop$ correspond à la moyenne des données de préférence ayant une note supérieur à JAR, - $MoyenneLikingPasAssez$ correspond à la moyenne des données de préférence ayant une note inférieur à JAR.

Calcul de la différence standardisée pour le niveau Trop

$$DifférenceStandardisée_{Trop} = \frac{EffetSurLaMoyenne_{Trop}}{\sqrt{(MCE(\frac{1}{n_{JAR}} + \frac{1}{n_{Trop}}))}}.$$

Calcul de la différence standardisée pour le niveau Pas assez

$$DifférenceStandardisée_{PasAssez} = \frac{EffetSurLaMoyenne_{PasAssez}}{\sqrt{(MCE(\frac{1}{n_{JAR}} + \frac{1}{n_{PasAssez}}))}}.$$

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.

OK : cliquez sur ce bouton pour lancer les calculs.

Annuler : cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

Aide : cliquez sur ce bouton pour afficher l'aide.

 : cliquez sur ce bouton pour rétablir les options par défaut.

 : cliquez sur ce bouton pour effacer les sélections de données.

Onglet **Général** :

Données de préférence : sélectionnez les données de préférence. Plusieurs colonnes peuvent éventuellement être sélectionnées. Si des en-têtes de colonnes ont été sélectionnés, veuillez vérifier que l'option « Libellés des colonnes » est activée.

Données sur l'échelle JAR : sélectionnez les données mesurées sur l'échelle JAR. Plusieurs colonnes peuvent être sélectionnées. Si des en-têtes de colonnes ont été sélectionnés, veuillez vérifier que l'option « Libellés des colonnes » est activée.

- **Echelle** : choisissez le type d'échelle (valeur par défaut : 1 -> 5).

Libellés des 3 niveaux JAR : activez cette option si vous voulez utiliser des libellés pour les 3 niveaux JAR. Cela peut vous permettre de rendre les résultats plus lisibles. Si l'option « Libellés des colonnes » est activée, la première cellule de la sélection doit comprendre un en-tête.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des colonnes : activez cette option si la première ligne des données sélectionnées (données de préférence, données sur l'échelle JAR, libellés des 3 niveaux JAR) contient un

libellé.

Poids : activez cette option si vous voulez pondérer les observations. Si vous n'activez pas cette option, les poids seront tous considérés comme valant 1. Les poids doivent être impérativement supérieurs ou égaux à 0. Si un en- tête de colonne a été sélectionné, veuillez vérifier que l'option « Libellés des colonnes » est activée.

Onglet **Options** :

Taille seuil pour la population : entrez le pourcentage de la population totale que doit représenter une catégorie pour être prise en compte dans les comparaisons multiples.

Onglet **Données manquantes** :

Ne pas accepter les valeurs manquantes : activez cette option pour que XLSTAT empêche la poursuite des calculs si des valeurs manquantes sont détectées.

Supprimer les observations : activez cette option pour supprimer les observations comportant des données manquantes.

Supprimer par colonne : activez cette option pour supprimer les données manquantes par colonne.

Estimer les données manquantes : activez cette option pour estimer les données manquantes avant le début des calculs.

- **Moyenne ou mode** : activez cette option pour estimer les données manquantes en utilisant la moyenne (variables quantitatives) ou le mode (variables qualitatives) pour les variables correspondantes.
- **Plus proche voisin** : activez cette option pour estimer les données manquantes d'une observation en recherchant le plus proche voisin de l'observation.

Onglet **Sorties** :

Statistiques descriptives : activez cette option pour afficher les statistiques descriptives pour l'ensemble des variables sélectionnées.

Corrélations : activez cette option pour afficher la matrice des corrélations des variables quantitatives sélectionnées. Si toutes les données sont ordinales, il est recommandé d'utiliser le coefficient de corrélation de Spearman.

Tableau à trois niveaux : activez cette option pour afficher le tableau des données JAR une fois effectué le regroupement des 5 catégories en 3 catégories.

Tableau des pénalités : activez cette option pour afficher le tableau présentant les impacts sur la moyenne ainsi que les pénalités.

Comparaisons multiples : activez cette option pour effectuer et afficher des comparaisons multiples de moyennes. Plusieurs méthodes de comparaison multiples sont proposées, regroupées en deux catégories : les comparaisons par paires, et les comparaisons à un groupe témoin, en l'occurrence le groupe JAR.

- **Niveau de signification (%)** : entrez le niveau de signification pour déterminer si les différences sont significatives ou non.

Onglet **Graphiques** :

Barres empilées : activez cette option pour afficher un graphique sous forme de barres empilées, permettant de visualiser les effectifs relatifs des différents groupes de l'échelle JAR.

- **3D** : activez cette option pour afficher des barres en trois dimensions.

Synthèse : activez cette option pour afficher les graphiques résumant les comparaisons multiples.

Effets sur la moyenne vs % : activez cette option pour afficher un graphique permettant de visualiser les effets sur les moyennes (pas assez, ou trop) en fonction du % de testeurs correspondant.

Résultats

Après l'affichage des statistiques simples pour l'ensemble des données sélectionnées (préférence et JAR), et de la matrice des corrélations correspondante, XLSTAT affiche un tableau présentant pour chacune des variables JAR les **effectifs pour les 5 niveaux** (pour le cas de l'échelle 1 à 5). Le diagramme en « barres empilées » correspondant est ensuite affiché.

Le tableau des données agrégées sur trois niveaux est ensuite affiché suivi du tableau des **effectifs agrégés sur 3 niveaux**. Le diagramme en « barres empilées » correspondant est ensuite affiché.

Le tableau des pénalités fourni ensuite les statistiques pour les 3 niveaux, y compris les moyennes, les impacts sur la moyenne, les pénalités, et les résultats des tests de comparaison.

Enfin les graphiques de synthèse permettent de rapidement identifier les caractéristiques JAR pour lesquelles les différences entre le groupe « JAR » et les groupes « 2 » et « 4 » sont significativement différentes : lorsque la différence est significative les barres sont affichées en rouge, alors qu'elles sont affichées en vert lorsque la différence n'est pas significative. Les barres apparaissent en gris lorsque l'effectif d'un groupe est inférieur au seuil choisi (voir l'onglet Options de la boîte de dialogue).

Le dernier graphique (effets sur la moyenne vs %) permet de visualiser les effets sur les moyennes (pas assez, ou trop) en fonction du % de testeurs correspondant. Le % de population seuil choisi pour considérer qu'un résultat est significatif est affiché sur la forme d'une ligne pointillée.

Exemple

Un exemple de *penalty analysis* est disponible sur le Centre d'aide XLSTAT à l'adresse

http://www.xlstat.com/demo-penf.htm

Bibliographie

Popper P., Schlich P., Delwiche J., Meullenet J.-F., Xiong R., Moskovitz H., Lesniauskas R.O., Carr T.B., Eberhardt K., Rossi F., Vigneau E. Qannari, Courcoux P. and Marketo C. (2004). Workshop summary : Data Analysis workshop : getting the most out of just-about-right data. *Food Quality and Preference*, 15, 891-899.

Analyse de données de Tri Libre

Utilisez cette fonction pour analyser des données de Tri libre rapidement et efficacement.

Cette fonction permet :

- d'étudier et visualiser les liens entres les produits
- d'étudier les accords entre les sujets.

Dans cette section :

Description

Les épreuves de Tri Libre (ou Free Sorting en Anglais) sont de plus en plus populaires dans le cadre de la caractérisation sensorielle des produits. Elles sont faciles à construire et il est facile d'y répondre. Le principe est le suivant : chaque participant doit faire des groupes avec l'ensemble des produits qui lui sont donnés. Chaque groupe représente ainsi un ensemble de produits se ressemblant fortement pour le sujet qui l'a construit. Le nombre de groupes est choisi par le participant lui-même. Les seules choses interdites sont de mettre tous les produits dans un même groupe et de faire autant de groupes que de produits.

Dans le cas où les libellés de vos groupes sont importants, vous pouvez utiliser la méthode ACM qui vous permet d'analyser les étiquettes.

Enfin, si vous avez un plan incomplet, vous pouvez utiliser des données manquantes. Cependant, uniquement la méthode AFC sur la matrice de cooccurence vous permettra d'analyser vos données. Attention, il est fortement préconisé que: * Chacun des produits soit vu le même nombre de fois * Chaque paire de produit soit vu le même nombre de fois * Le nombre de sujets soit conséquent

Méthodes contenues dans l'analyse de données de tri libre

L'objectif principal de cette analyse est de construire une représentation graphique des produits. Pour cela, 3 méthodes existent dans XLSTAT :

1. **STATIS** : Un prétraitement des données est réalisé permettant de pouvoir utiliser la méthode STATIS. Le pré-traitement consiste à considérer chaque sujet comme un tableau disjonctif complet, où les tailles des groupes sont ensuite standardisées (Llobell, Cariou, Vigneau, Labenne & Qannari, 2020). La méthode STATIS permet d'avoir des indices d'accord entre les sujets et de tenir compte de ces derniers dans l'analyse.

2. **AFC** sur la matrice de co-occurrence : une matrice de cooccurrence des produits est construite, suivie d'une Analyse Factorielle des Correspondances (Cariou & Qannari, 2018). Possède l'avantage de gérer les plans incomplets.

3. **ACM** : On effectue directement une Analyse des Correspondances Multiples sur les données (Van der Kloot & Van Herk, 1991). Possède l'avantage d'analyser les libellés des groupes.

Un autre objectif est de représenter les sujets. Pour cela, une matrice de cooccurrence des sujets est construite, suivie d'une Analyse Factorielle des Correspondances (Cariou & Qannari, 2018) (uniquement si le plan est complet).

Structure des données

Chaque ligne représente un produit et chaque colonne représente un sujet. Dans chaque colonne se trouvent simplement les numéros (ou noms) des groupes auxquels appartiennent les produits. Prenons un exemple où un sujet a formé un groupe avec les produits P1 et P3, et un groupe avec les produits P2 et P4. Il y aura alors un 1 (ou G1, "groupe 1", ...) pour P1 et P3 et un 2 (ou G2, "groupe 2", ...) pour P2 et P4. Si vous avez un plan incomplet, indiquez des valeurs manquantes quand le sujet n'a pas vu le produit.

Interprétation des résultats

La représentation des produits (resp. sujets) dans l'espace des k facteurs permet d'interpréter visuellement les proximités entre les produits (resp. sujets), moyennant certaines précautions.

On peut considérer que la projection d'un produit ou d'un sujet sur un plan est fiable si elle est éloignée du centre du graphique.

Nombre de facteurs

Deux méthodes sont communément utilisées pour déterminer le nombre de facteurs k à retenir pour l'interprétation des résultats :

- Regarder la courbe décroissante des valeurs propres. Le nombre de facteurs à retenir correspond au premier point d'inflexion sur la courbe.

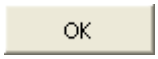
- On peut aussi se baser sur le pourcentage cumulé de variabilité représenté par les axes factoriels et décider de se contenter d'un certain pourcentage.

Représentations graphiques

Les représentations graphiques ne sont fiables que si la somme des pourcentages de variabilité associés aux axes de l'espace de représentation est suffisamment élevée. Si ce pourcentage est élevé (par exemple 80%), on peut considérer que la représentation est fiable. Si le pourcentage est faible, il est conseillé de faire des représentations sur plusieurs paires d'axes afin de valider l'interprétation faite sur les deux premiers axes factoriels.

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.


: cliquez sur ce bouton pour changer la façon dont vous chargez les données dans XLSTAT. Si le bouton possède une icône de souris, XLSTAT vous permet de faire une sélection à la souris des données. Si le bouton possède une icône de liste, XLSTAT vous permet de sélectionner les données à partir d'une liste. Si le bouton possède une icône orange, des boutons complémentaires avec un point d'interrogation  sont affichés pour vous permettre d'importer les données à partir de fichiers.

Onglet **Général** :

Données de Tri Libre : sélectionnez les données correspondant aux différents sujets. Si la première ligne de la sélection comprend des en-têtes, l'option « Libellés des sujets » doit être activée.

Méthode : pour représenter les produits, 3 méthodes sont possibles :

- **STATIS** : choisissez cette option si vous voulez utiliser la méthode STATIS suite à un pré-traitement adapté.
- **AFC sur la matrice de cooccurrence** : choisissez cette option si vous voulez utiliser une Analyse Factorielle des Correspondances sur la matrice de cooccurrence des produits.
- **ACM** : choisissez cette option si vous voulez utiliser une Analyse des Correspondances Multiples sur les données brutes.

Libellés des produits : activez cette option si vous voulez utiliser des libellés des produits pour l'affichage des résultats. Si l'option « Libellés des sujets » est activée, la première cellule de la sélection doit comprendre un en-tête. Si vous n'activez pas cette option, des libellés seront automatiquement créés.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des sujets : activez cette option si la première ligne des données sélectionnées (Données de Tri Libre, Libellés des produits) contient un libellé. Si vous n'activez pas cette option, des libellés seront automatiquement créés.

Onglet **Options** :

Filtrer les facteurs : vous pouvez activer l'une des deux options suivantes afin de réduire le nombre de facteurs pour lesquels les résultats sont affichés :

- **% minimum** : activez cette option puis saisissez le pourcentage minimum de la variabilité totale que doivent représenter les facteurs retenus.
- **Nombre maximum** : activez cette option pour fixer le nombre maximum de facteurs à prendre en compte.

Onglet **Sorties** :

L'onglet Sorties est divisé en plusieurs sous-onglets :

Général :

Ces sorties concernent toutes les méthodes :

Statistiques descriptives : activez cette option pour afficher les statistiques descriptives pour l'ensemble des sujets sélectionnés.

Analyse des sujets : activez cette option si vous voulez faire une analyse des sujets.

- **Matrice de cooccurrence** : activez cette option pour afficher la matrice de cooccurrence des sujets.
- **Valeurs propres de l'AFC** : activez cette option pour afficher le tableau des valeurs propres de l'AFC sur la matrice de cooccurrence des sujets.
- **Coordonnées des sujets** : activez cette option pour afficher les coordonnées des sujets dans l'espace des facteurs.

STATIS :

Ces sorties concernent uniquement l'analyse STATIS, et ne sont disponibles que si c'est la méthode que vous avez choisie :

Valeurs propres : activez cette option pour afficher le tableau des valeurs propres.

Coordonnées du consensus : activez cette option pour afficher les coordonnées du consensus dans l'espace des facteurs.

Matrice RV : activez cette option pour afficher la matrice des coefficients RV entre les sujets.

Facteurs de mise à l'échelle : activez cette option pour afficher les facteurs de mise à l'échelle des sujets.

Poids : activez cette option pour afficher les poids créés et utilisés par STATIS.

Configuration consensus : activez cette option pour afficher la configuration consensus créée par STATIS.

Homogénéité : activez cette option pour afficher l'homogénéité des sujets.

RV sujets/consensus : activez cette option pour afficher le coefficient RV entre chaque sujet et le consensus.

Erreur globale : activez cette option pour afficher l'erreur du critère STATIS.

Résidu par sujet : activez cette option pour afficher les résidus du critère STATIS pour chaque sujet.

Résidu par produit : activez cette option pour afficher les résidus du critère STATIS pour chaque produit.

AFC :

Ces sorties concernent uniquement l'AFC sur la matrice de cooccurrence des produits, et ne sont disponibles que si c'est la méthode que vous avez choisie :

Matrice de cooccurrence : activez cette option pour afficher la matrice de cooccurrence des produits.

Valeurs propres de l'AFC : activez cette option pour afficher le tableau des valeurs propres de l'AFC sur la matrice de cooccurrence.

Coordonnées des produits : activez cette option pour afficher les coordonnées des produits dans l'espace des facteurs.

ACM :

Ces sorties concernent uniquement l'ACM, et ne sont disponibles que si c'est la méthode que vous avez choisie :

Valeurs propres : activez cette option pour afficher le tableau des valeurs propres de l'ACM.

Coordonnées des produits : activez cette option pour afficher les coordonnées des produits dans l'espace des facteurs.

Contributions des produits : activez cette option pour afficher le tableau des contributions des produits.

Etiquettes : activez cette option pour afficher les coordonnées des groupes dans l'espace des facteurs.

Onglet **Graphiques** :

Général :

Ces sorties concernent toutes les analyses :

Graphiques sur 2 axes : activez cette option pour que XLSTAT ne vous demande pas de sélectionner les axes, et affiche automatiquement les 2 premiers axes.

Analyse des sujets :

- **Valeurs propres de l'AFC** : activez cette option pour afficher le graphique (*scree plot*) des valeurs propres de l'AFC sur la matrice de cooccurrence des sujets.
- **Coordonnées des sujets** : activez cette option pour afficher les graphiques des coordonnées des sujets dans l'espace des facteurs (en fonction du nombre de facteurs choisi).

STATIS :

Ces sorties concernent uniquement l'analyse STATIS, et ne sont disponibles que si c'est la méthode que vous avez choisie :

Valeurs propres : activez cette option pour afficher le graphique (*scree plot*) des valeurs propres.

Coordonnées du consensus : activez cette option pour afficher le graphique des coordonnées du consensus dans l'espace des facteurs.

Facteurs de mise à l'échelle : activez cette option pour afficher le diagramme en bâtons des facteurs de mise à l'échelle des sujets.

Poids : activez cette option pour afficher le diagramme en bâtons des poids créés et utilisés par STATIS.

RV sujets/consensus : activez cette option pour afficher le diagramme en bâtons des coefficient RV entre chaque sujet et le consensus.

Résidus par sujet : activez cette option pour afficher le diagramme en bâtons des résidus du critère STATIS pour chaque sujet.

Résidus par produit : activez cette option pour afficher le diagramme en bâtons des résidus du critère STATIS pour chaque produit.

Graphiques des nuages partiels : activez cette option pour afficher le graphique représentant à la fois les produits et les produits de chacun des sujets projeté dans l'espace des facteurs.

- Libellés des produits: activez cette option pour afficher les libellés des produits sur les graphiques.
- Libellés des nuages partiels : activez cette option pour afficher les libellés des points des nuages partiels.

AFC :

Ces sorties concernent uniquement l'AFC sur la matrice de cooccurrence des produits, et ne sont disponibles que si c'est la méthode que vous avez choisie :

- **Valeurs propres de l'AFC** : activez cette option pour afficher le graphique (*scree plot*) des valeurs propres de l'AFC sur la matrice de cooccurrence des produits.
- **Coordonnées des produits** : activez cette option pour afficher les graphiques des coordonnées des produits dans l'espace des facteurs (en fonction du nombre de facteurs choisi).

ACM :

Ces sorties concernent uniquement l'ACM, et ne sont disponibles que si c'est la méthode que vous avez choisie :

- **Valeurs propres** : activez cette option pour afficher le graphique (*scree plot*) des valeurs propres de l'ACM.
- **Coordonnées des produits** : activez cette option pour afficher les graphiques des coordonnées des produits dans l'espace des facteurs (en fonction du nombre de facteurs choisi).

Etiquettes : activez cette option pour afficher les graphiques des groupes dans l'espace des facteurs. Si les graphiques des coordonnées des produits sont aussi sélectionné, alors un biplot sera affiché.

Résultats

Statistiques descriptives : le tableau des statistiques descriptives présente des statistiques simples pour tous les sujets sélectionnés. Les groupes de chaque sujet avec leur taille respective et pourcentages de produits dans chaque groupe sont affichés.

STATIS :

Ces résultats concernent uniquement l'analyse STATIS, et ne sont disponibles que si c'est la méthode que vous avez choisie :

Valeurs propres et pourcentages d'inertie : les valeurs propres et le graphique (*scree plot*) correspondant sont affichés.

Coordonnées du consensus : les coordonnées du consensus dans l'espace des facteurs sont affichées, ainsi que les graphiques correspondants (en fonction du nombre de facteurs choisi).

Matrice RV : la matrice des coefficients RV entre tous les sujets est affichée. Le coefficient RV est un coefficient de similarité entre deux sujets compris entre 0 et 1. Plus il est proche de 1, plus la similarité est forte. Cette matrice est utilisée par STATIS pour calculer les poids des sujets.

Facteur d'échelle pour chaque sujet : les facteurs d'échelle sont affichés, ainsi que le diagramme en bâtons associé. Ces facteurs d'échelles permettent de standardiser le nombre de groupes de chaque sujet. Moins un sujet a fait de groupes, plus son facteur d'échelle est grand.

Poids de chaque sujet : les poids calculés par STATIS sont affichés, ainsi que le diagramme en bâtons associé. Plus un poids est grand, plus le sujet a contribué à l'élaboration du consensus. Sachant que STATIS donne du poids aux sujets les plus proches du point de vue global, un poids beaucoup plus faible que les autres signifiera que le sujet est atypique.

Configuration consensus : le consensus créé par STATIS est affiché. Il correspond à la somme des données sujets pré-traitées pondérée par les poids de ces sujets.

Homogénéité : l'homogénéité des sujets est affichée. C'est une valeur comprise entre $1/m$ (m étant le nombre de sujets) et 1, qui croît avec l'homogénéité des sujets.

Coefficient RV entre chaque sujet et le consensus : les coefficients RV entre les sujets et le consensus sont affichés, ainsi que le diagramme en bâtons associé. Tout comme les poids de STATIS, ces coefficients permettent de détecter des sujets atypiques. L'avantage de ces coefficients est qu'ils sont compris entre 0 et 1, donc plus faciles à interpréter que les poids.

Erreur globale : l'erreur du critère STATIS est affichée. Elle correspond à la somme de tous les résidus (qui peuvent être présentés par sujet ou par produit).

Résidus par sujet : ce tableau et le diagramme en bâtons correspondant permettent de visualiser la répartition des résidus de STATIS par sujet. On peut ainsi repérer quels sujets se démarquent le plus du consensus.

Résidus par produit : ce tableau et le diagramme en bâtons correspondant permettent de visualiser la répartition des résidus de STATIS par produit. On peut ainsi repérer pour quels produits STATIS a été moins efficace, autrement dit, quels produits se démarquent le plus de la configuration consensus.

Graphiques des nuages partiels : les nuages partiels correspondent aux projections des produits de chacun des sujets dans l'espace des facteurs. La représentation des points des nuages partiels superposée avec celles des produits permet de visualiser à la fois la diversité de l'information apportée par les différents sujets pour un produit donné, et de visualiser les distances relatives entre deux produits en fonction des différents sujets.

AFC sur la matrice de cooccurrence :

Ces résultats concernent uniquement l'AFC sur la matrice de cooccurrence des produits, et ne sont disponibles que si c'est la méthode que vous avez choisie :

Matrice de cooccurrence : la matrice de cooccurrence entre tous les produits est affichée. Cette matrice symétrique permet de voir combien de fois deux produits ont été placés dans le même groupes par les sujets.

Valeurs propres et pourcentages d'inertie : les valeurs propres de l'AFC et le graphique (*scree plot*) correspondant sont affichés.

Coordonnées des produits : les coordonnées des produits dans l'espace des facteurs sont affichées, ainsi que les graphiques correspondants (en fonction du nombre de facteurs choisi).

Analyse des Correspondances Multiples :

Ces résultats concernent uniquement l'ACM, et ne sont disponibles que si c'est la méthode que vous avez choisie :

Valeurs propres et pourcentages d'inertie : les valeurs propres de l'ACM et le graphique (*scree plot*) correspondant sont affichés.

Coordonnées des produits : les coordonnées des produits dans l'espace des facteurs sont affichées, ainsi que les graphiques correspondants (en fonction du nombre de facteurs choisi).

Contributions des produits : les contributions des produits sont affichées. Les contributions sont une aide à l'interprétation. Les produits ayant influencé le plus la construction des axes sont ceux dont les contributions sont les plus élevées.

Etiquettes : les contributions des groupes de chacun des sujets sont affichées. Un biplot est ensuite affiché.

Analyse des sujets :

Matrice de cooccurrence : la matrice de cooccurrence entre tous les sujets est affichée. Cette matrice symétrique permet de voir combien de fois deux sujets ont tous les deux placés dans le même groupe deux produits différents.

Valeurs propres et pourcentages d'inertie : les valeurs propres de l'AFC et le graphique (*scree plot*) correspondant sont affichés.

Coordonnées des sujets : les coordonnées des sujets dans l'espace des facteurs sont affichées, ainsi que les graphiques correspondants (en fonction du nombre de facteurs choisi).

Exemple

Un exemple d'utilisation d'Analyse de données de Tri Libre est disponible sur le Centre d'aide XLSTAT à l'adresse suivante :

http://www.xlstat.com/demo-fstf.htm

Bibliographie

Cariou, V., Qannari, E. M. (2018). Statistical treatment of free sorting data by means of correspondence and cluster analyses. *Food Quality and Preference*, **68**, 1-11.

Courcoux, P., Qannari, E. M., & Faye, P. (2015). Free sorting as a sensory profiling technique for product development. In Rapid Sensory Profiling Techniques (pp. 153-185). Woodhead Publishing.

Llobell, F., Cariou, V., Vigneau, E., Labenne, A., Qannari, E. M. (2020). Analysis and clustering of multiblock datasets by means of the STATIS and CLUSTATIS methods. Application to sensometrics. *Food Quality and Preference*, **79**, 103520.

Llobell, F. (2020). Classification de tableaux de données, applications en analyse sensorielle (Doctoral dissertation, Nantes, Ecole nationale vétérinaire).

Van der Kloot, W. A., & Van Herk, H. (1991). Multidimensional scaling of sorting data: A comparison of three procedures. *Multivariate Behavioral Research*, **26(4)**, 563-581.

Analyse de données de projective mapping

Utilisez cette fonction pour analyser des données de projective mapping rapidement et efficacement.

Cette fonction permet :

- d'étudier et visualiser les liens entre les produits
- d'étudier les accords entre les sujets.

Dans cette section :

Description

L'épreuve de projective mapping (ou Napping) est une des épreuves dites "rapides" de plus en plus populaires dans le cadre de la caractérisation sensorielle des produits. Elle consiste à demander à chacun des sujets de placer les produits sur une feuille de papier. Les données récupérées sont simplement les coordonnées des produits en abscisse et en ordonnée sur la feuille de papier. Chacun des sujets apporte donc un tableau à n lignes (une par produit) et 2 colonnes. Ces données peuvent être analysées avec la méthode STATIS ou avec l'Analyse Factorielle Multiple (AFM). Si les deux méthodes ont toutes les deux pour objectif principal de synthétiser l'information pour représenter graphiquement les produits, elles permettent aussi de déterminer les liens entre les réponses des sujets.

Structure des données

Chaque ligne représente un produit et les colonnes sont les coordonnées en abscisse et en ordonnée pour chacun des sujets. Les données des sujets sont juxtaposées verticalement. Si les données ne sont **pas adaptées** à ce type de tableau, il est possible de les **transformer** en un tableau horizontal, comme décrit précédemment, en utilisant la fonctionnalité **Créer un tableau Produits\Sujets**.

Interprétation des résultats

La représentation des produits dans l'espace des k facteurs permet d'interpréter visuellement les proximités entre les produits, moyennant certaines précautions.

On peut considérer que la projection d'un produit sur un plan est fiable si elle est éloignée du centre du graphique.

Nombre de facteurs

Deux méthodes sont communément utilisées pour déterminer le nombre de facteurs k à retenir pour l'interprétation des résultats :

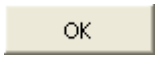
- Regarder la courbe décroissante des valeurs propres. Le nombre de facteurs à retenir correspond alors au premier point d'inflexion sur la courbe.
- On peut aussi se baser sur le pourcentage cumulé de variabilité représenté par les axes factoriels et décider de se contenter d'un certain pourcentage.

Représentations graphiques

Les représentations graphiques ne sont fiables que si la somme des pourcentages de variabilité associés aux axes de l'espace de représentation est suffisamment élevée. Si ce pourcentage est élevé (par exemple 80%), on peut considérer que la représentation est fiable. Si le pourcentage est faible, il est conseillé de faire des représentations sur plusieurs paires d'axes afin de valider l'interprétation faite sur les deux premiers axes factoriels.

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.


: cliquez sur ce bouton pour changer la façon dont vous chargez les données dans XLSTAT. Si le bouton possède une icône de souris, XLSTAT vous permet de faire une sélection à la souris des données. Si le bouton possède une icône de liste, XLSTAT vous permet de sélectionner les données à partir d'une liste. Si le bouton possède une icône orange, des

boutons complémentaires avec un point d'interrogation  sont affichés pour vous permettre d'importer les données à partir de fichiers.

Onglet **Général** :

Données de projective mapping : sélectionnez les données correspondant aux différents sujets. Si la première ligne de la sélection comprend des en-têtes, l'option « Libellés des coordonnées » doit être activée. Si les données ne sont **pas adaptées** à ce type de tableau, il est possible de les **transformer** en un tableau horizontal, comme décrit précédemment, en utilisant la fonctionnalité **Créer un tableau Produits\Sujets.**

Méthode : pour représenter les produits, deux méthodes sont possibles :

- **STATIS** : choisissez cette option si vous voulez utiliser la méthode STATIS.
- **AFM** : choisissez cette option si vous voulez réaliser une Analyse Factorielle Multiple.

Libellés des produits : activez cette option si vous voulez utiliser des libellés des produits pour l'affichage des résultats. Si l'option « Libellés des coordonnées » est activée, la première cellule de la sélection doit comprendre un en-tête. Si vous n'activez pas cette option, des libellés seront automatiquement créés.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des coordonnées : activez cette option si la première ligne des données sélectionnées (Données de projective mapping, Libellés des produits, Libellés des sujets) contient un libellé. Si vous n'activez pas cette option, des libellés seront automatiquement créés.

Libellés des sujets : activez cette option si vous voulez utiliser des libellés des sujets pour l'affichage des résultats. Le nombre de libellés doit être identique au nombre de sujets dans les données de projective mapping. Si l'option « Libellés des coordonnées » est activée, la première cellule de la sélection doit comprendre un en-tête. Si vous n'activez pas cette option, des libellés seront automatiquement créés.

Onglet **Options** :

Filtrer les facteurs : vous pouvez activer l'une des deux options suivantes afin de réduire le nombre de facteurs pour lesquels les résultats sont affichés :

- **% minimum** : activez cette option puis saisissez le pourcentage minimum de la variabilité totale que doivent représenter les facteurs retenus.

- **Nombre maximum** : activez cette option pour fixer le nombre maximum de facteurs à prendre en compte.

Graphiques sur les deux premiers axes : activez cette option pour que XLSTAT ne vous demande pas de sélectionner les axes, et affiche automatiquement les deux premiers axes.

Onglet **Données manquantes** :

Ne pas accepter les valeurs manquantes : activez cette option pour que XLSTAT empêche la poursuite des calculs si des valeurs manquantes sont détectées.

Supprimer les observations : activez cette option pour supprimer les observations comportant des données manquantes.

Estimer les données manquantes : activez cette option pour estimer les données manquantes avant le début des calculs.

- **Moyenne ou mode** : activez cette option pour estimer les données manquantes en utilisant la moyenne (variables quantitatives) ou le mode (variables qualitatives) pour les variables correspondantes.
- **Plus proche voisin** : activez cette option pour estimer les données manquantes d'une observation en recherchant le plus proche voisin de l'observation.

Onglet **Sorties** :

L'onglet Sorties est divisé en plusieurs sous-onglets :

STATIS :

Ces sorties concernent uniquement l'analyse STATIS, et ne sont disponibles que si c'est la méthode que vous avez choisie :

Statistiques descriptives : activez cette option pour afficher les statistiques descriptives pour l'ensemble des sujets sélectionnés.

Valeurs propres : activez cette option pour afficher le tableau des valeurs propres.

Coordonnées du consensus : activez cette option pour afficher les coordonnées du consensus dans l'espace des facteurs.

Matrice RV : activez cette option pour afficher la matrice des coefficients RV entre les sujets.

Facteurs de mise à l'échelle : activez cette option pour afficher les facteurs de mise à l'échelle des sujets.

Poids : activez cette option pour afficher les poids créés et utilisés par STATIS.

Configuration consensus : activez cette option pour afficher la configuration consensus créée par STATIS.

Homogénéité : activez cette option pour afficher l'homogénéité des sujets.

RV sujets/consensus : activez cette option pour afficher le coefficient RV entre chaque sujet et le consensus.

Erreur globale : activez cette option pour afficher l'erreur du critère STATIS.

Résidu par sujet : activez cette option pour afficher les résidus du critère STATIS pour chaque sujet.

Résidu par produit : activez cette option pour afficher les résidus du critère STATIS pour chaque produit.

AFM :

Ces sorties concernent uniquement l'AFM, et ne sont disponibles que si c'est la méthode que vous avez choisie :

Statistiques descriptives : activez cette option pour afficher les statistiques descriptives pour l'ensemble des sujets sélectionnés.

Valeurs propres : activez cette option pour afficher le tableau des valeurs propres de l'AFM.

Coordonnées des produits : activez cette option pour afficher les coordonnées des produits dans l'espace des facteurs.

Contributions des produits : activez cette option pour afficher le tableau des contributions des produits.

Cosinus carrés : activez cette option pour afficher le tableau des cosinus carrés des produits.

Coefficients Lg : activez cette option pour afficher les coefficients Lg de liaison entre les sujets.

Onglet **Graphiques** :

STATIS :

Ces graphiques concernent uniquement l'analyse STATIS, et ne sont disponibles que si c'est la méthode que vous avez choisie :

Valeurs propres : activez cette option pour afficher le graphique (*scree plot*) des valeurs propres.

Coordonnées du consensus : activez cette option pour afficher le graphique des coordonnées du consensus dans l'espace des facteurs.

Facteurs de mise à l'échelle : activez cette option pour afficher le diagramme en bâtons des facteurs de mise à l'échelle des sujets.

Poids : activez cette option pour afficher le diagramme en bâtons des poids créés et utilisés par STATIS.

RV sujets/consensus : activez cette option pour afficher le diagramme en bâtons des coefficients RV entre chaque sujet et le consensus.

Résidus par sujet : activez cette option pour afficher le diagramme en bâtons des résidus du critère STATIS pour chaque sujet.

Résidus par produit : activez cette option pour afficher le diagramme en bâtons des résidus du critère STATIS pour chaque produit.

Graphiques des nuages partiels : activez cette option pour afficher le graphique représentant à la fois les produits du consensus et les produits de chacun des sujets projetés dans l'espace des facteurs.

- **Libellés des produits** : activez cette option pour afficher les libellés des produits sur les graphiques.
- **Libellés des nuages partiels** : activez cette option pour afficher les libellés des points des nuages partiels.

AFM :

Ces graphiques concernent uniquement l'AFM, et ne sont disponibles que si c'est la méthode que vous avez choisie :

- **Valeurs propres** : activez cette option pour afficher le graphique (*scree plot*) des valeurs propres de l'AFM.
- **Coordonnées des produits** : activez cette option pour afficher les graphiques des coordonnées des produits dans l'espace des facteurs (en fonction du nombre de facteurs choisi).

Graphiques des nuages partiels : activez cette option pour afficher le graphique représentant à la fois les produits du consensus et les produits de chacun des sujets projetés dans l'espace des facteurs.

- **Libellés des produits** : activez cette option pour afficher les libellés des produits sur les graphiques.
- **Libellés des nuages partiels** : activez cette option pour afficher les libellés des points des nuages partiels.

Résultats

Statistiques descriptives : le tableau des statistiques descriptives présente des statistiques simples pour tous les sujets sélectionnés.

STATIS :

Ces résultats concernent uniquement l'analyse STATIS, et ne sont disponibles que si c'est la méthode que vous avez choisie :

Valeurs propres et pourcentages d'inertie : les valeurs propres et le graphique (*scree plot*) correspondant sont affichés.

Coordonnées du consensus : les coordonnées du consensus dans l'espace des facteurs sont affichées, ainsi que les graphiques correspondants (en fonction du nombre de facteurs choisi).

Matrice RV : la matrice des coefficients RV entre tous les sujets est affichée. Le coefficient RV est un coefficient de similarité entre deux sujets compris entre 0 et 1. Plus il est proche de 1, plus la similarité est forte. Cette matrice est utilisée par STATIS pour calculer les poids des sujets.

Facteur d'échelle pour chaque sujet : les facteurs d'échelle sont affichés, ainsi que le diagramme en bâtons associé. Ces facteurs d'échelles permettent de standardiser l'utilisation de la feuille de chaque sujet. Moins un sujet a mis d'espace entre ses produits, plus son facteur d'échelle est grand.

Poids de chaque sujet : les poids calculés par STATIS sont affichés, ainsi que le diagramme en bâtons associé. Plus un poids est grand, plus le sujet a contribué à l'élaboration du consensus. Sachant que STATIS donne du poids aux sujets les plus proches du point de vue global, un poids beaucoup plus faible que les autres signifiera que le sujet est atypique.

Configuration consensus : le consensus créé par STATIS est affiché. Il correspond à la somme des données sujets pré-traitées pondérée par les poids de ces sujets.

Homogénéité : l'homogénéité des sujets est affichée. C'est une valeur comprise entre $1/m$ (m étant le nombre de sujets) et 1, qui croît avec l'homogénéité des sujets.

Coefficient RV entre chaque sujet et le consensus : les coefficients RV entre les sujets et le consensus sont affichés, ainsi que le diagramme en bâtons associé. Tout comme les poids de STATIS, ces coefficients permettent de détecter des sujets atypiques. L'avantage de ces coefficients est qu'ils sont compris entre 0 et 1, donc plus faciles à interpréter que les poids.

Erreur globale : l'erreur du critère STATIS est affichée. Elle correspond à la somme de tous les résidus (qui peuvent être présentés par sujet ou par produit).

Résidus par sujet : ce tableau et le diagramme en bâtons correspondant permettent de visualiser la répartition des résidus de STATIS par sujet. On peut ainsi repérer quels sujets se démarquent le plus du consensus.

Résidus par produit : ce tableau et le diagramme en bâtons correspondant permettent de visualiser la répartition des résidus de STATIS par produit. On peut ainsi repérer pour quels produits STATIS a été moins efficace, autrement dit, quels produits se démarquent le plus de la configuration consensus d'un sujet à l'autre.

Graphiques des nuages partiels : les nuages partiels correspondent aux projections des produits de chacun des sujets dans l'espace des facteurs. La représentation des points des nuages partiels superposée avec celles des produits permet de visualiser à la fois la diversité de l'information apportée par les différents sujets pour un produit donné, et de visualiser les distances relatives entre deux produits en fonction des différents sujets.

AFM :

Ces résultats concernent uniquement l'AFM, et ne sont disponibles que si c'est la méthode que vous avez choisie :

Valeurs propres et pourcentages d'inertie : les valeurs propres de l'AFM et le graphique (*scree plot*) correspondant sont affichés.

Coordonnées des produits : les coordonnées des produits dans l'espace des facteurs sont affichées, ainsi que les graphiques correspondants (en fonction du nombre de facteurs choisi).

Contributions des produits : les contributions des produits sont affichées. Les contributions sont une aide à l'interprétation. Les produits ayant influencé le plus la construction des axes sont ceux dont les contributions sont les plus élevées.

Cosinus carrés : On peut considérer que la projection d'un point sur un axe, un plan ou un espace à 3 dimensions est fiable si la somme des cosinus carrés sur les axes de représentation n'est pas trop éloignée de 1. Les cosinus carrés sont affichés dans les résultats proposés par XLSTAT afin d'éviter toute mauvaise interprétation.

Coefficients Lg : les coefficients Lg de liaison entre les sujets permettent de mesurer à quel point les sujets sont proches deux à deux.

Graphiques des nuages partiels : les nuages partiels correspondent aux projections des produits de chacun des sujets dans l'espace des facteurs. La représentation des points des nuages partiels superposée avec celles des produits permet de visualiser à la fois la diversité de l'information apportée par les différents sujets pour un produit donné, et de visualiser les distances relatives entre deux produits en fonction des différents sujets.

Exemple

Un exemple d'utilisation d'analyse de données de projective mapping est disponible sur le Centre d'aide XLSTAT à l'adresse suivante :

http://www.xlstat.com/demo-prmf.htm

Bibliographie

Llobell, F. (2020). Classification de tableaux de données, applications en analyse sensorielle (Doctoral dissertation, Nantes, Ecole nationale vétérinaire).

Llobell, F., Cariou, V., Vigneau, E., Labenne, A., Qannari, E. M. (2020). Analysis and clustering of multiblock datasets by means of the STATIS and CLUSTATIS methods. Application to sensometrics. *Food Quality and Preference*, **79**, 103520.

Pagès, J. (2005). Collection and analysis of perceived product inter-distances using multiple factor analysis: Application to the study of 10 white wines from the Loire Valley. *Food Quality and Preference*, **16(7)**, 642–649.

Risvik, E., McEwan, J. A., & Rødbotten, M. (1997). Evaluation of sensory profiling and projective mapping data. *Food Quality and Preference*, **8(1)**, 63–71.

Analyse de données CATA

Utilisez cette fonction pour analyser des données CATA (*check-all-that-apply*) rapidement et efficacement. Si l'enquête CATA comprend des données de préférence, cet outil peut être utilisé pour identifier des facteurs de satisfaction ou, au contraire, des caractéristiques considérées négatives par les consommateurs.

Dans cette section :

Description

Depuis 2007, lorsqu'elles ont été présentées par Adams *et al*., les enquêtes CATA (check-all-that-apply) sont de plus en plus populaires dans le cadre de la caractérisation sensorielle des produits. Les enquêtes CATA s'adressent aux consommateurs, plus représentatifs du marché, plutôt qu'à des sujets entraînés. Elles sont faciles à construire et il est facile d'y répondre. Le principe est le suivant : chaque participant (sujet) reçoit un questionnaire contenant des attributs ou descripteurs appliqués à un ou plusieurs produits. Pour chaque produit, le sujet coche les attributs qui selon lui s'applique au produit ou ne les coche pas dans le cas contraire. D'autres questions (utilisant des échelles différentes) peuvent être ajoutées pour relier les attributs à des scores de préférence. Si les participants doivent donner une note globale à chaque produit impliqué dans l'étude, des analyses plus poussées telles que la modélisation de préférence sont envisageables. Afin d'améliorer la reproductibilité, Ares *et al*. (2014) recommandent de randomiser l'ordre des questions CATA pour chaque participant.

L'outil d'analyse de données CATA de XLSTAT a été développé dans le but d'automatiser cette analyse. Des améliorations ont été apportées par l'équipe d'XLSTAT en 2020 afin de permettre de mieux évaluer la qualité des données avant l'analyse.

Considérons une enquête menée sur N sujets pour P produits (un des produits pouvant être virtuel, souvent idéal) décrits par K attributs. Les données CATA pour les K attributs sont enregistrées sous forme binaire (1 pour coché, 0 pour non coché). Trois formats de données sont acceptés par XLSTAT :

1. Format Horizontal (P x (N x K)) : XLSTAT s'attend à un tableau Excel avec P lignes et N groupes de K colonnes, les groupes étant placés côte à côte. L'utilisateur spécifie la valeur de N et XLSTAT déduit automatiquement la valeur de K. Si l'enquête implique une question sur la préférence, la colonne correspondante peut être introduite au sein de chaque groupe de K colonnes à une position qui peut être indiquée à XLSTAT. Dans ce cas, chaque groupe sera

associé à K+1 colonnes. Si l'un des produits correspond à un produit idéal, il est possible d'indiquer sa position.

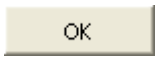
2. Format Horizontal (N x (P x K)) : XLSTAT s'attend à un tableau Excel avec N lignes et P groupes de K colonnes, les groupes étant placés côte à côte. L'utilisateur spécifie la valeur de P et XLSTAT déduit automatiquement la valeur de K. Si l'enquête implique une question sur la préférence, la colonne correspondante peut être introduite au sein de chaque groupe de K colonnes à une position qui peut être indiquée à XLSTAT. Dans ce cas, chaque groupe sera associé à K+1 colonnes. Si l'un des produits correspond à un produit idéal, il est possible d'indiquer sa position.

3. Format Vertical ((N x P) x K) : XLSTAT s'attend à un tableau Excel avec P x N lignes et K colonnes. Les identifiants des produits et ceux des sujets doivent être saisis dans deux champs supplémentaires. Si l'enquête implique une question sur la préférence, la colonne correspondante doit être sélectionnée. Si un des produits correspond à un produit idéal, vous pouvez indiquer son identifiant afin qu'XLSTAT le considère comme tel.

Les analyses effectuées par XLSTAT sur des données CATA sont basées sur l'article de Meyners *et al*. (2013), qui explore en profondeur les possibilités offertes par les données CATA.

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.

Onglet **Général** :

Données CATA (0/1) : sélectionnez les données CATA (0/1).

Format : sélectionnez le format des données CATA saisies. Ce format peut être **horizontal** ou **vertical** (pour plus de détails, voir partie description). Si le libellé des variables a été sélectionné, veuillez vérifier que l'option « Libellés inclus » est activée.

Si le format est **horizontal** :

- **(P X K X N)**

Nombre de sujets : entrez le nombre de sujets (N). XLSTAT déduira le nombre de descripteurs (K).

- **(N x K x P)**

Nombre de produits : entrez le nombre de produits (P). XLSTAT déduira le nombre de descripteurs (K).

Position du produit idéal : indiquez si le produit idéal se trouve à une certaine position dans le tableau CATA ou s'il se trouve en dernière position.

Données de préférence : indiquez si les données de préférence se trouvent à une certaine position dans le tableau CATA ou si elles se trouvent en dernière position. Il faut une colonne préférence par sujet et une valeur pour chaque produit. Cette valeur peut être manquante pour le produit idéal.

Libellés des produits : activez cette option si des libellés de produits sont disponibles, puis sélectionnez les données correspondantes. Si le libellé des variables a été sélectionné, veuillez vérifier que l'option « Libellés inclus » est activée.

Libellés des sujets : activez cette option si des libellés de sujets sont disponibles, puis sélectionnez les données correspondantes. Si le libellé des variables a été sélectionné, veuillez vérifier que l'option « Libellés inclus » est activée.

Si le format est **vertical ((N x P) x K)** :

Produits : sélectionnez les données correspondant aux produits testés. Une colonne unique doit être sélectionnée. Si le libellé des variables a été sélectionné, veuillez vérifier que l'option « Libellés inclus » est activée.

Sujets : sélectionnez les données correspondant aux identifiants des sujets. Une colonne unique doit être sélectionnée. Si le libellé des variables a été sélectionné, veuillez vérifier que l'option « Libellés inclus » est activée.

Données de préférence : activez cette option si des données de préférence sont disponibles, puis sélectionnez les données correspondantes. Une colonne unique doit être sélectionnée. Si le libellé des variables a été sélectionné, veuillez vérifier que l'option « Libellés inclus » est activée.

Produit idéal : activez cette option si les sujets ont qualifié un produit idéal, et spécifiez l'identifiant de ce produit parmi les identifiants des produits.

Plage : activez cette option pour afficher les résultats à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Onglet **Options (1)** :

Validation des données CATA : Activez cette option pour que XLSTAT valide la qualité des données CATA.

Test Q de Cochran : activez cette option pour effectuer des tests de Cochran sur chacun des attributs.

- **Comparaisons multiples par paires** : choisissez la méthode de comparaisons multiples à utiliser. Deux méthodes sont proposées : Différence critique ou McNemar (Bonferroni) pour réaliser des tests de McNemar avec un seuil modifié. Pour plus de détails sur ces options veuillez vous reporter au test de Cochran.

- **Différences (ou p-values) par attribut** : activez cette option pour afficher pour chaque attribut le tableau des différences (option Différence critique) ou des p-values (option McNemar Bonferroni) entre les paires de produits. Les valeurs en gras représentent les différences significatives.

- **Filtrer les attributs non significatifs** : activez cette option pour supprimer les attributs pour lesquels les testsQ de Cochran ne sont pas significatifs pour un seuil que vous pouvez choisir.

Indépendance des attributs : activez cette option pour tester l'indépendance des attributs au travers d'un test du Khi2 multivarié. Si le test rejette l'hypothèse nulle d'indépendance, alors des comparaisons multiples permettent d'identifier quelles paire d'attributs sont potentiellement liées.

Analyse factorielle des correspondances :

Distance : sélectionnez la distance sur laquelle baser l'analyse factorielle des correspondances : khi-deux pour l'analyse factorielle des correspondances (AFC) classique, ou Hellinger dans le cas où certains attributs sont peu cités.

Test d'indépendance : activez cette option pour effectuer un test d'indépendance sur la table de contingence.

Niveau de signification (%) : entrer le niveau de signification pour le test. Cette valeur est également utilisée pour déterminer la significativité des tests Q de Cochran.

Filtrer les facteurs : vous pouvez activer une des deux options suivantes afin de réduire le nombre de facteurs à afficher :

- **% Minimum** : activez cette option puis entrez le pourcentage minimal à atteindre pour déterminer le nombre de facteurs à afficher.

- **Nombre maximum** : activez cette option pour paramétrer le nombre maximal de facteurs à prendre en compte dans l'affichage des résultats.

Onglet **Options (2)** :

Filtrer les produits : activez cette option pour pouvoir choisir quels produits sont pris en compte dans l'analyse CATA.

Filtrer les sujets : activez cette option pour pouvoir choisir quels sujets sont pris en compte dans l'analyse CATA.

Taille seuil pour la population : entrez le pourcentage de la population totale que doit représenter une catégorie pour être prise en compte dans l'analyse des effets sur la moyenne au travers de l'analyse des pénalités.

Dans le cas d'un produit idéal présent, ce pourcentage sera le minimum nécessaire de l'effectif du tableau de contingence produit idéal/produits.

Dans le cas où il n'y a pas de produit idéal, les catégories seront présent et absent.

Gestion des sessions multiples : dans le cas où les sujets auraient noté plusieurs fois un produit, veuillez indiquer comme XLSTAT doit se comporter.

- **Arrêter les calculs** : XLSTAT interrompt les calculs.
- **Agréger les sessions (priorité aux 1)** : XLSTAT agrège les triplets sujet/produit/attributs en un seul, en prenant 1 si l'attribut a été coché une fois et 0 sinon.

Onglet **Données manquantes** :

Ne pas accepter les valeurs manquantes : activez cette option pour que XLSTAT empêche la poursuite des calculs si des valeurs manquantes sont détectées.

Remplacer les données manquantes par 0 : activez cette option si vous considérez que les données manquantes correspondent à des 0.

Résultats

Validation des données CATA

Si l'option correspondante a été cochée dans la boîte de dialogue, XLSTAT affiche dans un premier temps une série de résultats qui permettent de contrôler la qualité des données CATA. Les données sont analysées d'abord pour les sujets pour identifier d'éventuelles anomalies, puis ensuite pour identifier des anomalies au niveau des attributs.

Test de Cochran

Des tests Q de Cochran sont effectués sur le tableau Sujets x Produits, pour chaque attribut indépendamment. La première colonne du tableau fournit pour chaque attribut (en lignes) la p-value. Puis des comparaisons multiples basées sur l'approche de McNemar Bonferroni ou de Marascuilo sont effectuées. Les colonnes qui suivent fournissent la proportion de sujets ayant coché le produit pour l'attribut en question. Les lettres entre parenthèses sont à considérer uniquement si la p-value est significative. Elles peuvent être utilisées pour identifier les produits responsables du rejet de l'hypothèse nulle d'égalité des produits. Le test Q de Cochran est l'équivalent d'un test de McNemar dans le cas où il n'y aurait que deux produits.

Test d'indépendance des attributs

Un test du Khi2 multivarié est effectué pour tester si les attributs sont indépendants les uns des autres. Si l'hypothèse d'indépendance est rejetée, des tests de comparaison par paires (tests de Fisher exacts) sont effectués afin d'identifier quels attributs sont liés.

Analyse Factorielle des Correspondances

Les données CATA sont résumées dans un tableau de contingence (somme des N tables CATA individuelles ; la valeur maximale de chaque cellule est N). Une Analyse Factorielle des Correspondances (AFC) est réalisée afin de visualiser le tableau de contingence. L'AFC peut être basée sur la distance du Khi² ou sur la distance de Hellinger (aussi connue sous le nom de distance de Bhattacharya, nommée ainsi dans l'outil matrices de similarité/dissimilarité de XLSTAT). La distance de Hellinger entre deux échantillons ne dépend que de la distribution de ces deux échantillons. L'analyse factorielle des correspondances basée sur la distance de Hellinger est donc adaptée au cas où certains attributs sont peu sélectionnés. Les attributs associés à une somme marginale nulle sont éliminés de l'analyse factorielle des correspondances. Les résultats suivants sont affichés : tableau de contingence, test d'indépendance entre les lignes et les colonnes, valeurs propres et pourcentage d'inertie, et graphique symétrique ou asymétrique (respectivement pour les options Khi² et Hellinger).

Analyse en Coordonnées Principale

Les corrélations tétrachoriques (adaptées aux données binaires) entre descripteurs et, si des données de préférence sont présentes, les corrélations bisérielles (développées pour mesurer la corrélation entre une variable binaire et une variable quantitative) entre chaque descripteur et les données de préférence sont calculées et visualisées par une Analyse en Coordonnées Principales (PCOA) basée sur la correction de Lingoes lorsque c'est nécessaire. Les valeurs propres et les coordonnées principales ainsi que leur représentation graphique sont affichés. Les proximités entre descripteurs peuvent être analysées.

Analyses des pénalités

Si des données de préférence sont disponibles, des analyses des pénalités sont effectuées. Si un produit idéal a été évalué, deux analyses sont effectuées, pour les attributs nécessaires (P(No)|(Yes) et P(Yes)|(Yes)) et pour les attributs positifs (P(Yes)|(No) and P(No)|(No)). Si aucun produit idéal n'a été évalué, l'analyse des pénalités est effectuée sur les présences / absences des attributs.

Un tableau de synthèse contient les fréquences d'apparition des deux situations (P(No)|(Yes) et P(Yes)|(Yes) ou P(Yes)|(No) et P(No)|(No) ou présence et absence) pour chaque attribut.

Le tableau de comparaison contient les effets sur la moyenne des scores de préférence entre les deux situations et leurs significativités. Le graphique de représentation des effets sur la moyenne et le graphique des effets sur la moyenne vs % illustrent le tableau précédent. Si un produit idéal est présent, l'analyse des attributs nécessaires et l'analyses des attributs positifs sont résumées sur un graphique de synthèse des effets sur la moyenne vs %.

Analyse des attributs

Une série de K tableaux 2x2 (un tableau par attribut) est affichée, avec en lignes, les valeurs enregistrées pour le produit idéal et en colonnes, les valeurs obtenues pour les produits testés. Les cellules du tableau contiennent les préférences moyennes (moyennées sur les sujets et les produits) et le % de tous les cas associés à la combinaison correspondante de 0s et/ou de 1s.

XLSTAT considère deux produits comparables si la valeur absolue de leur différence est inférieure à 1.

Produit idéal\Produits	0	1
0	6.2 (12%)	7.4 (8%)
1	5.1 (39%)	7.2 (41%)

Pour un attribut donné,

- si l'attribut est coché pour le produit idéal (seconde ligne), et si la préférence pour les produits cochés (cellule [1,1]) est supérieure à la préférence pour les produits non cochés (cellule [1,0]), alors l'attribut est « **nécessaire** ».
- Symétriquement, si l'attribut n'est pas coché pour le produit idéal (première ligne) et si la préférence pour les produits non cochés (cellule [0,0]) est supérieure à la préférence pour les produits cochés (cellule [0,1]), alors l'attribut est « **négatif** ».
- Si la (cellule [0,1]) > (cellule [0,0]) significativement, alors l'attribut est **intéressant**.
- Si l'attribut n'est pas coché pour le produit idéal (première ligne) et si la préférence pour les produits cochés (cellule [0,1]) est comparable à celle pour les produits non cochés (cellule [0,0]) alors l'attribut est « **indifférent** ».
- Enfin, si l'attribut n'est pas nécessaire et que la préférence pour les produits cochés (cellule [1,1]) est comparable à celle pour les produits non cochés (cellule [1,0]), l'attribut est **sans influence**.

Certains tableaux peuvent correspondre aux trois situations. XLSTAT associera chaque tableau à une situation, mais vous pouvez contrôler les résultats. XLSTAT tentera de relier chaque tableau 2x2 à une des règles définies plus haut, dans le même ordre.

Lors d'une épreuve CATA contenant un produit idéal, l'objectif pour les produits est de se rapprocher au plus près de ce produit idéal. Par conséquent, le graphique représentant la différence de citation (nombre de cochages) entre le produit idéal et le produit en question s'avère très utile. Pour chaque attribut, nous pouvons voir si le produit est semblable ou différent du produit idéal. Plus un attribut est sujet à des différences, plus il est problématique et se situera à gauche du graphique. À l'inverse, plus pour un attribut donné le produit est semblable au produit idéal, plus la ligne sera proche de 0. Si la différence est négative, l'attribut n'est pas assez présent, alors que si elle positive, il est trop présent. Pour terminer, l'intervalle de confiance permet de déterminer si la différence avec le produit idéal est significative.

Exemple

Un exemple d'analyse de données CATA est disponible sur le Centre d'aide XLSTAT à l'adresse

http://www.xlstat.com/demo-catadataf.htm

Bibliographie

Ares G., Antúnez L., Roigard C.M., Pineau B. Hunter D. and Jaeger S. (2014). Further investigations into the reproducibility of check-all-that- apply (CATA) questions for sensory product characterization elicited by Consumers. *Food Quality and Preference*, **36**, 111-121.

Cuadras C. M. & Cuadras i Pallejà D. (2008). A unified approach for representing rows and columns in contingency tables.

Meyners M., Castura J. C. and Carr B. T. (2013). Existing and new approaches for the analysis of CATA data. *Food Quality and Preference*, **30**, 309-319.

Analyse de données TCATA

Utilisez la méthode TCATA (Temporal Check-All-That-Apply) pour analyser vos données TCATA. Cette méthode permet aux évaluateurs de sélectionner et de mettre à jour en continu les attributs qui caractérisent les produits au fil du temps.

Dans cette section :

Description

La méthode TCATA est une extension temporelle de la méthode CATA (Check-All-That-Apply) développée par Castura *et al* (2016). Cette méthode permet de décrire les propriétés sensorielles multidimensionnelles des produits au fur et à mesure de leur évolution au cours du temps. La sélection et la désélection des attributs sont suivies en continu au fil du temps, permettant aux évaluateurs de caractériser l'évolution des changements sensoriels des produits. Des résultats graphiques permettent de visualiser l'évolution des profils sensoriels au cours du temps et de comparer les produits.

Les données TCATA doivent absolument être équilibrées, c'est-à-dire que chaque sujet doit évaluer chaque produit dans chaque session. Deux formats de données sont acceptés pour les données TCATA :

- **Binaire** : Les données comportent autant de lignes qu'il y a de combinaisons produits, sujets, attributs et éventuellement sessions, et autant de colonnes que de points temps. Les données sont enregistrées sous forme binaire, 1 si l'attribut est sélectionné 0 sinon.
- **Temps de début/fin** : Deux colonnes sont attendues. Pour chaque ligne correspondant à une combinaison produit/sujet/attribut et éventuellement session, la colonne « Début » contient le temps de départ où l'attribut a été cochée, la colonne « Fin » contient le temps où l'attribut a été décoché. Si le même attribut est resélectionné plus tard, alors une seconde ligne avec la même combinaison produit sujet attribut est nécessaire.

Interprétation des résultats

Les principaux résultats spécifiques à la méthode TCATA sont les suivants :

- **Courbes des proportions de citation** : Pour chaque produit on affiche les courbes de proportion de citation de chacun des attributs en fonction du temps. Une option permet de lisser les courbes. On peut également afficher une courbe de référence, pour chaque attribut. Pour un produit donné, cette courbe de référence correspond à la proportion moyenne de citation pour l'ensemble des autres produits rassemblés. Afin de tester la

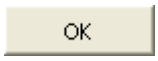
différence entre la courbe d'un attribut et sa courbe de référence on réalise un test de Fisher ou du Khi². Afin de ne pas surcharger le graphique, les courbes de référence sont affichées uniquement si la différence est significative pour un temps donné. Lorsqu'une courbe de référence significative est affichée la courbe de l'attribut concerné est affichée en gras.

- **Différences entre produits** : Un graphique est affiché pour chaque paire de produits. Pour chaque attribut on regarde si la différence de proportions de citation est significative ou pas à l'aide d'un test de Fisher ou du Khi². Lorsque la différence est significative pour un temps donné, on affiche la courbe de différence de proportions.

- **Trajectoire des produits** : Une analyse factorielle des correspondances est réalisée afin de définir la trajectoire des produits. Chaque ligne correspond à une combinaison produit/temps et les colonnes contiennent les attributs. Les coordonnées factorielles des lignes (produit/temps) sont reliées entre elles et ainsi les trajectoires des produits au fil du temps sont affichées sur le plan factoriel. Le point final (temps maximum) de chaque trajectoire contient l'étiquette avec le nom du produit.

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.


: cliquez sur ce bouton pour changer la façon dont vous chargez les données dans XLSTAT. Si le bouton possède une icône de souris, XLSTAT vous permet de faire une sélection à la souris des données. Si le bouton possède une icône de liste, XLSTAT vous permet de sélectionner les données à partir d'une liste. Si le bouton possède une icône orange, des boutons complémentaires avec un point d'interrogation  sont affichés pour vous permettre d'importer les données à partir de fichiers.

Onglet **Général**:

Format : sélectionnez le format des données TCATA saisies. Ce format peut être **Binaire** ou **Temps de début/fin** (pour plus de détails, voir la partie description).

Format binaire :

Données TCATA(0/1) : sélectionnez les données binaires TCATA. Si l'option libellés des variables est activée, la première ligne de la sélection sera utilisée en tant que valeurs temporelles. Dans le cas contraire les valeurs temporelles seront créées à partir de 0 et jusqu'au nombre de colonnes de la sélection.

Format Temps de début/fin :

Début : sélectionnez les données correspondant au temps de début de citation pour chaque combinaison sujet/produit/attribut. Si l'option « Libellés des variables » est activée, la première cellule de la sélection doit comprendre un en-tête. **Fin** : sélectionnez les données correspondant au temps de fin de citation pour chaque combinaison sujet/produit/attribut. Si l'option « Libellés des variables » est activée, la première cellule de la sélection doit comprendre un en-tête.

Produits : sélectionnez les données qui correspondent aux produits testés. Cette sélection ne peut contenir qu'une seule colonne. Si des en-têtes de colonnes ont été sélectionnés, veuillez vérifier que l'option « Libellés des variables » est activée.

Sujets : sélectionnez les données qui correspondent aux sujets. Cette sélection ne peut contenir qu'une seule colonne. Si des en-têtes de colonnes ont été sélectionnés, veuillez vérifier que l'option « Libellés des variables » est activée.

Attributs : sélectionnez les données qui correspondent aux attributs. Cette sélection ne peut contenir qu'une seule colonne. Si des en-têtes de colonnes ont été sélectionnés, veuillez vérifier que l'option « Libellés des variables » est activée.

Sessions : activez cette option si plusieurs sessions de test ont eu lieu. Si tel est le cas, sélectionnez les données qui correspondent à la session. Cette sélection ne peut contenir qu'une seule colonne. Si des en-têtes de colonnes ont été sélectionnés, veuillez vérifier que l'option « Libellés des variables » est activée.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des variables : activez cette option si la première ligne des données sélectionnées contient un libellé.

Précision temporelle : cette option est uniquement disponible pour le format « Temps de début/fin ». Choisissez le niveau de précision à utiliser pour les données temporelles : 1 seconde ou 0.5 seconde.

Onglet **Données manquantes** :

Ne pas accepter les valeurs manquantes : activez cette option pour que XLSTAT empêche la poursuite des calculs si des valeurs manquantes sont détectées.

Remplacer les valeurs manquantes par 0 : activez cette option si vous considérez que les valeurs manquantes sont équivalentes à des 0.

Onglet **Sorties**:

AFC : activez cette option pour réaliser une AFC et éventuellement afficher la trajectoire des produits (pour plus de détails, voir la partie description).

- **Valeurs propres** : activez cette option pour afficher le tableau des valeurs propres de l'AFC.
- **Coordonnées des lignes** : activez cette option pour afficher les coordonnées des lignes du tableau sur lequel a été réalisé l'AFC, chaque ligne correspond à une combinaison produit/temps.
- **Coordonnées des colonnes** : activez cette option pour afficher les coordonnées des colonnes du tableau sur lequel a été réalisé l'AFC, chaque colonne correspond à un attribut.

Accord des sujets : activez cette option pour afficher le tableau contenant l'accord des sujets. Plus l'accord d'un sujet est proche de 1 plus sa manière d'évaluer les produits est proche de l'ensemble des autres sujets, plus son accord est proche de 0, plus sa manière d'évaluer les produits est différente de l'ensemble des autres sujets.

Répétabilité des sujets : si vous avez sélectionné plusieurs sessions, activez cette option pour afficher le tableau contenant la répétabilité des sujets. Si la répétabilité d'un sujet est proche de 1, cela indique qu'il évalue un même produit de la même manière entre les différentes sessions, à l'inverse, si sa répétabilité est proche de 0, cela indique qu'il a jugé les produits différemment entre les différentes sessions.

Si vous avez choisi de réaliser une AFC, vous pouvez filtrer le nombre de facteurs à afficher dans les tableaux des coordonnées des lignes et des colonnes :

- **% minimum** : activez cette option puis saisissez le pourcentage minimum de la variabilité totale que doivent représenter les facteurs retenus.

- **Nombre maximum** : activez cette option pour fixer le nombre maximum de facteurs à prendre en compte.

Onglet **Graphiques** :

Proportions de citation :

- **Chaque attribut par produit** : activez cette option pour afficher un diagramme en bâtons des proportions de citation. Pour chaque combinaison produit/attribut, un graphique sera affiché.
- **Courbes par produit** : activez cette option pour rassembler sur un même graphique associé à un produit les courbes de proportions de citation de tous les attributs. Autant de graphiques que de produits seront affichés. Vous pouvez choisir d'afficher les courbes de référence relatives aux attributs (pour plus de détails, voir la partie description).
    - **Courbe brute** : activez cette option pour afficher les courbes brutes, c'est à dire avec un point pour chaque pas de temps.
    - **Courbe lissée** : activez cette option pour affichers des courbes lissés.

Différence entre produits : activez cette option pour afficher pour chaque paire de produits les différences significatives de proportions de citation (pour plus de détails, voir la partie description).

Graphique de l'AFC : si vous avez choisi dans les options de réaliser une AFC, vous pouvez choisir d'afficher le graphique correspondant à la trajectoire des produits (pour plus de détails, voir la partie description).

- **Trajectoires des produits** : activez cette option pour afficher la trajectoire des produits sur les plans factoriels de l'AFC.
    - **Graphiques sur deux axes** : activez cette option pour afficher les trajectoires des produits que sur les deux premiers axes.
    - **Colorer les produits** : activez cette option pour colorer chaque produit différemment.

Si vous avez choisi d'afficher les courbes lissées, différents paramètres de lissage vous sont proposés :

Nombre de nœuds :

- **Automatique** : activez cette option pour calculer automatiquement le nombre ainsi que la localisation des nœuds (points réels) de la courbe lissée. Moins de nœuds seront retenus pour les parties de la courbe ayant une petite courbure alors qu'un plus grand nombre de nœuds sera retenu pour les parties fortement incurvées.
- **Défini par l'utilisateur** : activez cette option pour définir manuellement le nombre de nœuds de la courbe. Les coordonnées de ces derniers seront uniformément réparties sur l'axe des abscisses.

- **Tolérance** : entrez le niveau de tolérance de lissage à appliquer sur les courbes (la valeur vaut 0.001 par défaut).

Type de test : L'affichage des courbes de référence ou des différences entre produits nécessite le calcul de tests statistiques (pour plus de détails, voir la partie description). Le test de Fisher est celui utilisé par Castura *et al* (2016), cependant si vous avez beaucoup de produits ou d'attributs le grand nombre de tests à réaliser peut ralentir grandement le temps de calcul. C'est pourquoi XLSTAT propose également de réaliser un test du Khi² qui à l'avantage d'être beaucoup plus rapide.

- **Niveau de signification** : entrez le niveau de signification pour le test sélectionné.

Résultats

Tableau de synthèse : un tableau de synthèse contenant le nombre de sessions, de sujets, de produits et d'attributs est affiché. Les temps de début et de fin d'évaluation sont également affichés. Si vos données ne sont pas équilibrées (pour plus de détails, voir la partie description) un tableau indiquant les combinaisons manquantes est affiché.

Diagramme en bâtons des proportions de citation pour chaque attribut de chaque produit : si vous avez activé l'option correspondante, les diagrammes en bâtons des proportions de citations sont affichés pour chaque combinaison produit/attribut.

Courbes par produit : si vous avez activé l'option correspondante, les courbes des proportions de citation de chaque attribut sont affichées pour chaque produit.

Différences significatives de proportions de citation entre produits : si vous avez activé l'option correspondante, les courbes de différences significatives de proportion sont affichées pour chaque paire de produits.

Accord des sujets : si vous avez activé l'option correspondante, le tableau contenant les accords des sujets est affiché.

Répétabilité des sujets : si vous avez activé l'option correspondante, le tableau contenant les répétabilités des sujets est affiché.

Analyse factorielle des correspondances (AFC) : si vous avez choisi de calculer l'AFC, les résultats seront affichés. Le graphique des trajectoires des produits est également affiché.

Exemple

Un exemple d'utilisation de la méthode TCATA est disponible sur le Centre d'aide XLSTAT à l'adresse suivante : http://www.xlstat.com/demo-tcaf.htm

Bibliographie

Castura, J.C., Antúnez, L., Giménez, A., & Ares, G. (2016). Temporal Check-All-That-Apply (TCATA): A novel dynamic method for characterizing products. *Food Quality and Preference*, **47**, 79–90.

Dominance temporelle des sensations

Utilisez cette fonction pour analyser des données de Dominance Temporelle des Sensations et identifier les descripteurs dominants au cours du temps pour des produits.

Dans cette section :

Description

La Dominance Temporelle des Sensations (TDS) est une méthode temporelle multidimensionnelle (Pineau, Cordelle, & Schlich, 2003). Une liste de descripteurs est présentée aux panelistes à qui il est demandé de choisir les attributs qu'ils jugent dominants au cours de la dégustation du produit. Une sensation dominante est celle qui attire le plus l'attention à un temps donné, mais pas forcément la sensation la plus intense à ce même moment. (Pineau *et al.*, 2009).

L'outils DTS d'XLSTAT permet d'analyser les attributs dominants pour un ensemble de produits.

Deux formats de données sont acceptés :

1. Format Dominance (0/1) : XLSTAT s'attend à un tableau Excel avec autant de lignes qu'il y a de combinaisons produits, panelistes, attributs et éventuellement sessions, et autant de colonnes que de points temps. Les données sont enregistrées sous forme binaire, 1 si l'attribut est sélectionné comme dominant, 0 sinon. Pour chaque combinaison paneliste*produit*session, chaque attribut ne peut apparaître qu'une fois.

2. Format Citations : XLSTAT s'attend à un tableau avec une ligne par attribut sélectionné par un paneliste pour un produit donnée et éventuellement une session donnée, et une colonne contenant les temps de sélection. Pour ce format de données, il est nécessaire d'avoir un attribut de début et un attribut de fin qui définisse les bornes de l'évaluation. Pour chaque combinaison paneliste*produit*session, un attribut apparaît autant de fois qu'il a été sélectionné comme dominant.

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.

OK : cliquez sur ce bouton pour lancer les calculs.

Annuler : cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

Aide : cliquez sur ce bouton pour afficher l'aide.

 : cliquez sur ce bouton pour rétablir les options par défaut.

 : cliquez sur ce bouton pour effacer les sélections de données.

Onglet **Général** :

Données DTS : sélectionnez les données DTS.

Format : sélectionnez le format des données DTS saisies. Ce format peut être **Dominance (0/1)** ou **Citations** (pour plus de détails, voir partie description). Si le libellé des variables a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

Produits : sélectionnez les données qui correspondent aux produits testés. Cette sélection ne peut contenir qu'une seule colonne. Si des en-têtes de colonnes ont été sélectionnés, veuillez vérifier que l'option « Libellés des variables » est activée.

Sujets : sélectionnez les données qui correspondent aux sujets. Cette sélection ne peut contenir qu'une seule colonne. Si des en-têtes de colonnes ont été sélectionnés, veuillez vérifier que l'option « Libellés des variables » est activée.

Attributs : sélectionnez les données qui correspondent aux attributs. Cette sélection ne peut contenir qu'une seule colonne. Si des en-têtes de colonnes ont été sélectionnés, veuillez vérifier que l'option « Libellés des variables » est activée.

Sessions : activez cette option si plusieurs sessions de test ont eu lieu. Si tel est le cas, sélectionnez les données qui correspondent à la session. Cette sélection ne peut contenir qu'une seule colonne. Si des en-têtes de colonnes ont été sélectionnés, veuillez vérifier que l'option « Libellés des variables » est activée.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des variables : activez cette option si la première ligne des données sélectionnées contient un libellé.

Onglet **Options** :

Standardisation du temps :

- **Aucune** : activez cette option pour garder les données telles qu'elles ont été saisies (pas de standardisation).
- **A droite** : activez cette option pour standardiser le temps afin de ramener toutes les fin d'évaluation à la même échelle. Les temps standardisés sont ramenés entre 0 et 1. Pour chaque paneliste*produit*session, les temps sont divisés par le temps maximum, c'est à dire le temps de fin d'évaluation.
- **A droite et à gauche** : activez cette option pour standardiser le temps afin de ramener tous les débuts et toutes les fins d'évaluation à la même échelle. Les temps standardisés sont ramenés entre 0 et 1. Pour chaque triplet paneliste*produit*session, les temps sont diminués du temps de début d'évaluation, c'est-à-dire du temps de sélection du premier attribut, et divisés par le temps maximum, c'est à dire le temps de fin d'évaluation.

Dominance :

- **Niveau de signification (%)** : entrez le niveau de signification pour le test.

Tolérance du lissage

- **Tolérance** : entrez le niveau de tolérance de lissage à appliquer sur les courbes DTS des produits (la valeur vaut 0.001 par défaut).

Onglet **Données manquantes** :

Ne pas accepter les valeurs manquantes : activez cette option pour qu'XLSTAT empêche la poursuite des calculs si des valeurs manquantes sont détectées.

Remplacer les données manquantes par 0 : activez cette option si vous considérez que les données manquantes correspondent à des 0 (seulement pour le format Dominance (0/1)).

Supprimer les observations : activez cette option pour supprimer les observations comportant des données manquantes.

Onglet **Graphiques** :

Courbes DTS : activez cette option pour afficher les courbes DTS des produits.

- **Lissage** : activez cette option pour lisser les courbes.
- **Automatique** : activez cette option pour calculer automatiquement le nombre ainsi que la localisation des noeuds (points réels) de la courbe lissée. Moins de noeuds seront retenus pour les parties de la courbe ayant une petite courbure alors qu'un plus grand nombre de noeuds sera retenu pour les parties fortement incurvées (voir dans la section *Options* pour définir le niveau approprié de tolérance de lissage).

- **Défini par l'utilisateur** : activez cette option pour définir manuellement le nombre de noeuds de la courbe. Les coordonnées de ces derniers seront uniformément réparties sur l'axe des abscisses.

- **Limite de chance** : activez cette option pour afficher la limite de chance du taux de dominance. La limite de chance est le taux de dominance qu'un attribut peut obtenir par chance. Elle est définie par $P_0 = \frac{1}{K}$, avec K le nombre d'attributs.

- **Limite de signification** : activez cette option pour afficher la limite de signification du taux de dominance. La limite de signification est la valeur minimum à atteindre par le taux de dominance pour qu'il soit significativement supérieur à P_0. Elle est calculée en utilisant l'intervalle de confiance d'une proportion binomiale basée sur une approximation Normale :

- $P_S = P_0 + z_\alpha \sqrt{\frac{P_0(1-P_0)}{J*S}}$, avec J le nombre de panelistes et S le nombre de sessions.

Bandes DTS : activez cette option pour afficher les bandes DTS des produits.

- **Bandes Oui/Non** : activez cette option pour afficher les attributs significativement dominants en une bande unique.

- **Bandes par attribut** : activez cette option pour afficher un graphique à 2 dimensions. Pour chaque période de dominance d'un attribut, une bande de hauteur relative au taux moyen de dominance sur cette période est tracée.

Résultats

Dominance par produit et par attribut : Le tableau de dominance par produit et par attribut contient les taux de dominance par produit et par attribut à chaque temps.

Courbes DTS : Pour chaque produit, un graphique avec les taux de dominance de chaque attribut en fonction du temps est affiché. Les taux de dominance sont définis pour chaque temps comme la proportion d'évaluations (paneliste*session) pour lesquelles l'attribut a été cité comme dominant (Pineau *et al.*, 2009). Si l'utilisateur le souhaite, les taux de dominance sont lissés en utilisant les splines cubiques. Si l'utilisateur le souhaite, les limites de chance et de signification sont tracées.

Bandes Oui/Non : Pour chaque produit, un graphique représentant une large bande contenant les attributs dominants à chaque temps est affiché. La bande est composée de rectangles colorés empilés (Monterymard *et al.*, 2010). La hauteur totale de la bande est constante. L'axe des abscisses représente le temps.

Bandes par attribut : Pour chaque produit, un graphique représentant chaque attribut dominant individuellement par des bandes est affiché. L'axe des abscisses représente le temps et l'axe des ordonnées représente les différents attributs. La hauteur des bandes est proportionnelle au taux moyen de dominance, permettant ainsi à l'utilisateur d'estimer l'importance de chaque attribut (Galmarini *et al.*, 2016). Pour chaque période de dominance, la hauteur de la bande est définie comme le taux de dominance moyen sur cette période divisé par le taux de dominance maximum pour ce produit.

Exemple

Un exemple d'analyse de données DTS est disponible sur le Centre d'aide XLSTAT à l'adresse

http://www.xlstat.com/demo-tdsf.htm

Bibliographie

Galmarini, M. V., Visalli, M., & Schlich, P. (2016). Advances in representation and analysis of mono and multi-intake Temporal Dominance of Sensations data. *Food Quality and Preference*.

Monterymard, C., Visalli, M., & Schlich, P. (2010).The TDS-band plot: A new graphical tool for temporal dominance of sensations data. In *2nd conference of the society of sensory professionals* (pp. 27–29).

Pineau, N., Cordelle, S., & Schlich, P. (2003). Temporal dominance of sensations: A new technique to record several sensory attributes simultaneously over time. In *5th Pangborn symposium* (p. 121).

Pineau, N., Schlich, P., Cordelle, S., Mathonnière, C., Issanchou, S., Imbert, A., Rogeaux, M., Etiévant, P. and Köster, E. (2009). Temporal Dominance of Sensations: Construction of the TDS curves and comparison with time–intensity. *Food Quality and Preference*, 20(6), pp.450-455.

Temps-Intensité

Utilisez cet outil pour analyser des données temps-intensité (TI) et identifier le profil temporel d'une sensation dans un ensemble de produits.

Dans cette section :

Description

Temps-Intensité est une méthode sensorielle temporelle qui a été introduite la première fois dans les années 30. Elle commence à être utilisée dans les années 50 (Söjström, 1954) et a réellement émergé dans les années 70 avec l'amélioration des appareils d'enregistrement.

Au cours d'une évaluation TI, les sujets doivent noter l'intensité de perception d'un unique attribut au cours de la consommation du produit. Comparé à des mesures points uniques, l'analyse du développement et du déclin d'une caractéristique sensorielle particulière peut révéler des informations riches pour distinguer des produits ou des perceptions. Ce type d'analyse peut s'appliquer à une grande variété de produits, allant du niveau de goût sucré d'une boisson, jusqu'à la sensation laissée par un tube à lèvres.

Les données TI consistent habituellement en plusieurs mesures d'intensité notée par un sujet et enregistrée à plusieurs intervalles de temps. Chacune de ces mesures doit être associée à un identificateur produit. XLSTAT offre également la possibilité d'indiquer un sujet ou bien un identificateur de session.

La première étape de l'analyse TI dans XLSTAT est de mesurer les paramètres caractéristiques des courbes sur chaque courbe individuelle. Le temps initial d'exposition au stimulus est considéré comme étant le premier point d'enregistrement de chaque courbe et il y a 10 paramètres distincts définis comme suit :

- I max : intensité pic ou intensité maximale observée sur toute la courbe ;
- T start : position en temps où la réaction au stimulus est perçue pour la première fois, définie comme la première valeur d'intensité excédant $X\%$ du pic d'intensité ;
- T max : position en temps du pic d'intensité sur la courbe ;

- T plateau : durée en temps autour de T max où l'intensité mesurée est plus grande que $(100 - X)\%$ du pic d'intensité ;
- T ext : position en temps de l'extinction de la perception du stimulus, définie comme le temps du premier point inférieur à $X\%$ de la valeur du pic d'intensité après le maximum d'intensité ;
- R croissant : pente ou taux d'accroissement de l'intensité entre T start et T max ;
- R décroissant : pente ou taux de décroissance de l'intensité entre T max et T ext ;
- Surface avant : l'aire sous la courbe avant le pic d'intensité ;
- Surface après : l'aire sous la courbe après le pic d'intensité ;
- Surface : l'aire totale sous la courbe, égale à la somme de la surface avant et après le pic.

Avec X la valeur du niveau de signification exprimé en $\%$.

Les paramètres de courbes mesurés sont affichés dans un tableau récapitulatif. Il est attendu que les courbes TI présentent une forme de cloche. Si pour une raison ou une autre, l'algorithme détecte que l'une ou plusieurs des courbes présente des caractéristiques pathologiques (intensité constante, plusieurs maximums, etc. …), un message est affiché de sorte que l'utilisateur peut investiguer quelle courbe devrait être retirée de l'analyse.

Le contrôle visuel de chaque courbe est une étape importante dans une analyse TI. L'utilisateur devrait utiliser son expertise de terrain pour s'assurer que les courbes possèdent les caractéristiques attendues. Pour cela, XLSTAT offre la possibilité d'afficher toutes les courbes enregistrées soit sur un graphique individuel, soit superposées sur un même graphique pour faciliter la comparaison entre les courbes.

En plus des courbes individuelles temps intensité, il est également très utile de résumer visuellement la perception du panel d'un stimulus pour plusieurs produits. Ceci peut être fait facilement avec XLSTAT en créant une courbe synthétique soit pour l'ensemble du jeu de données, soit pour chacun des produits. Plusieurs techniques sont proposées pour générer une courbe synthétique :

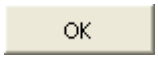
- Moyenne : la courbe synthétique est la moyenne par pas de temps de toutes les courbes TI individuelles ;
- Paramétrage ; la courbe synthétique est construite à partir des paramètres de courbes mesurés ;
- Méthode d'Overbosch : la courbe synthétique est créée en suivant l'approche proposée par Overbosch (1986) ;
- Méthode de Liu et MacFie : la courbe synthétique est créée en suivant l'approche proposée dans Liu (1990).

Les deux dernières techniques nécessitent la spécification d'un paramètre supplémentaire : le nombre de bins pour générer la courbe synthétique avant et après le pic d'intensité. Le nombre total de bins de la courbe synthétique est donc le double de ce chiffre.

Finalement, une ANOVA est exécutée sur chaque paramètre de courbe séparément dans le but d'évaluer l'effet produit et, optionnellement, les effets sujet et session. En fonction des effets sélectionnés, plusieurs configurations de modèles sont disponibles pour prendre en compte des interactions potentielles entre les produits, les sujets et les sessions. XLSTAT permet par ailleurs de considérer les sujets et / ou les sessions comme des effets fixes ou aléatoires.

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.

Onglet **Général** :

Données temps-intensité : sélectionnez les données temps-intensité. Ce sont les courbes telles qu'elles ont été enregistrée durant l'évaluation. Dans le mode de sélection par défaut d'XLSTAT (les colonnes sont des variables), il est attendu que les courbes soient organisées par lignes. Les colonnes correspondent alors aux pas de temps successifs. Si des en-têtes de colonnes ont été sélectionnés, veuillez vérifier que l'option « Libellés des variables » est activée.

Produits : sélectionnez les données qui correspondent aux produits testés. Cette sélection ne peut contenir qu'une seule colonne. Si des en-têtes de colonnes ont été sélectionnés, veuillez vérifier que l'option « Libellés des variables » est activée.

Sujets : activez cette option si plusieurs sujets ont évalué les produits. Sélectionnez les données qui correspondent aux sujets. Cette sélection ne peut contenir qu'une seule colonne. Si des en-têtes de colonnes ont été sélectionnés, veuillez vérifier que l'option « Libellés des variables » est activée.

Sessions : activez cette option si plusieurs sessions de test ont eu lieu. Si tel est le cas, sélectionnez les données qui correspondent à la session. Cette sélection ne peut contenir qu'une seule colonne. Si des en-têtes de colonnes ont été sélectionnés, veuillez vérifier que l'option « Libellés des variables » est activée.

Temps : activez cette option si vous souhaitez sélectionnez un vecteur de temps. Comme pour les données temps-intensité, il est attendu du vecteur de temps corresponde à une ligne dans le mode de sélection par défaut. Cette sélection ne peut contenir qu'une seule ligne, le vecteur temps est le même pour toutes les courbes individuelles temps intensité. Les colonnes correspondent alors aux pas de temps successifs. Si des en-têtes de colonnes ont été sélectionnés, veuillez vérifier que l'option « Libellés des variables » est activée.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des variables : activez cette option si la première ligne des données sélectionnées (variables dépendantes et explicatives, libellés des observations) contient un libellé.

Onglet **Options** :

Intervalle de confiance (%) : entrez l'intervalle de confiance utilisé pour déterminer en-deçà de quelle p-value, les différents tests amènent à rejeter l'hypothèse nulle associée.

Modèle : sélectionnez le modèle d'analyse de la variance (ANOVA) que vous souhaitez utilisé. Selon les effets qui ont été activés sur l'onglet général, plusieurs configurations de modèles sont proposées pour rendre compte des possibles interactions entre produits, sujets et sessions.

Effets aléatoires (Sujet / Session) : activez cette option si vous voulez considérer que les effets Sujet et Session et les éventuelles interactions les impliquant, soient considérés comme des effets aléatoires. Si cette option n'est pas activée, tous les effets sont considérés comme fixes.

Créer une courbe synthétique : activez cette option si vous souhaitez créer une courbe synthétique, utile pour récapituler visuellement la perception du panel en réponse à un stimulus sur différents produits (voir la description ci-dessus pour plus de détails). Les différentes techniques de calculs proposées dans XLSTAT sont :

- **Moyenne**
- **Paramétrage**
- **Méthode d'Ovecbosch**
- **Méthode de Liu et MacFie**

Une courbe par produit : activez cette option si vous souhaitez créer une courbe synthétique par produit.

Nombre de bins : si l'option **Créer une courbe synthétique** est activée et si, la méthode d'Overbosch ou de Liu et MacFie a été sélectionnée alors l'utilisateur doit saisir un nombre de bins. Ce nombre de bins est le nombre de données de la courbe synthétique avant et après le pic d'intensité de sorte que la courbe complète contienne deux fois ce nombre de points.

Onglet **Données manquantes** :

Ne pas accepter les données manquantes : activez cette option pour que les calculs soient stoppés lorsqu'une valeur manquante est détectée.

Estimer les données manquantes : activez cette option pour estimer les données manquantes avant le début des calculs.

- **Moyenne ou mode** : activez cette option pour estimer les données manquantes en utilisant la moyenne (variables quantitatives) ou le mode (variables qualitatives) pour les variables correspondantes.
- **Plus proche voisin** : activez cette option pour estimer les données manquantes d'une observation en recherchant le plus proche voisin de l'observation.

Onglet **Sorties** :

Statistiques descriptives : activez cette option pour afficher les statistiques descriptives pour les variables sélectionnées.

Paramètres des courbes : activez cette option pour afficher un tableau contenant les paramètres de courbes mesurés.

Tableau des courbes synthétiques : activez cette option pour afficher le tableau contenant les données des courbes synthétiques.

ANOVA : activez cette option pour afficher les tableaux de synthèse des différentes ANOVA effectuées.

Coefficients d'ajustement : dans ce tableau est affichée une série de statistiques pour le modèle.

Test de type III des effets fixes : activez cette option pour afficher le tableau de l'analyse de type III de la variance.

Effet produit : activez cette option pour afficher un tableau résumant l'effet produit pour chaque paramètre de courbe.

Interprétation : activez cette option pour afficher une aide additionnelle pour interpréter les résultats affichés.

Onglet **Graphiques** :

Courbes temps-intensité : activez cette option pour afficher les courbes individuelles temps-intensités sur des graphiques.

Afficher sur un seul graphique : activez cette option pour afficher toutes les courbes sur un même graphique.

Courbe synthétique : activez cette option pour afficher les courbes synthétiques sur des graphiques.

p-valeurs effet produit : activez cette option pour afficher un graphique résumé sur l'effet produit pour chaque paramètre de courbe.

Résultats

Statistiques descriptives : le tableau des statistiques descriptives montre des statistiques basiques sur les produits, sujets et sessions si activés. Le nom des différentes catégories sont affichés avec leurs fréquences respectives d'apparence.

Paramètres des courbes mesurés : le tableau récapitulatif des paramètres de courbes mesurés est affiché avec les produits, sujets et sessions associés.

Courbes synthétiques : Le tableau affiche les points des courbes synthétiques qui ont été générées (voir description pour plus de détails).

Ensuite pour chaque paramètre, une série de résultats d'ANOVA est affiché avec l'objectif de vérifier s'il y a un effet produit ou non. Pour chaque paramètre, le tableau des SS de type III de l'ANOVA est affiché pour le modèle sélectionné (voir description pour plus des détails). Enfin un tableau résumé compare les p-valeurs obtenues pour chaque paramètre.

Exemple

Un exemple d'utilisation de Temps Intensité est disponible sur le Centre d'aide XLSTAT à l'adresse

http://www.xlstat.com/demo-tif.htm

Bibliographie

Liu Y. H. and MacFie, H.J.H. (1990). "Methods for averaging Time- Intensity Curves", *Chemical Senses*, Vol. 15, 1990, pp. 471-484.

Overbosch P. *et al* . (1986). « An Improved Method for Measuring Perceived Intensity/Time Relationships in Human Taste and Smell ». Chemical Senses, Vol. 11, 1986, pp. 331-338.

Analyse sensorielle de durée de vie (shelf life analysis)

Cet outil permet d'étudier la durée de vie d'un produit en utilisant l'avis de sujets suite à des tests sensoriels. Il permet de trouver la période de consommation optimale. Pour se faire, XLSTAT utilise les modèles de survie paramétriques.

Dans cette section :

Description

L'analyse sensorielle de durée de vie d'un produit (sensorial shelf life analysis) permet d'évaluer la période idéale de consommation d'un produit en utilisant l'évaluation sensorielle de sujets à des temps différents.

Il peut arriver que les propriétés physico-chimiques d'un produit ne suffisent pas pour évaluer la qualité d'un produit en fonction de la période à laquelle il est consommé. On ajoute fréquemment une évaluation sensorielle du produit qui mettra en avant des périodes auxquelles le produit est optimal. Dans l'exemple d'un produit lacté, on pourra avoir un produit tout à fait adapté à la consommation mais qui, dans une évaluation sensorielle, sera trop acide après une certaine période ou aura un aspect moins attrayant.

L'utilisation de méthodes classiquement utilisées en analyse de données de survie s'applique très bien dans ce cas.

Généralement, lorsqu'on mène ce type de tests sensoriels, on fait goûter le même produit à des sujets à des périodes différentes. Ceci peut se faire lors de différentes sessions mais ce qui est généralement recommandé est de préparer un protocole qui permet d'obtenir des produits ayant des anciennetés différentes pour le jour du test.

Chaque sujet va exprimer son opinion sur le produit testé (aime / n'aime pas) et on obtiendra ainsi un tableau récapitulatif des opinions par temps et par sujet.

XLSTAT-MX permet d'avoir deux types de tableaux en entrée :

- Un tableau sujet $\times$ date : chaque colonne représente une date, chaque ligne représente un sujet. On aura deux valeurs différentes en fonction de l'appréciation du sujet (aime / aime pas).

- Une colonne de date et une colonne associée au nom des sujets. Pour chaque sujet, on entre la date où l'on observe qu'il a changé d'avis. En supposant que tous les sujets apprécient le produit lors de la première dégustation.

XLSTAT-MX utilise ensuite un modèle de survie paramétrique afin d'estimer un modèle de durée de vie du produit.

Comme les dates exactes auxquelles le sujet a changé d'avis ne sont pas connues, on utilise la notion de censure pour définir ces dates. Ainsi, si on a des relevés chaque semaine, si un sujet n'apprécie plus un produit après 3 semaines, on dira qu'il y a une censure par intervalle entre la $2^{ème}$ et la $3^{ème}$ semaine pour ce sujet. On suppose toujours qu'un sujet apprécie le produit lors de la première période testée. Si le sujet apprécie le produit durant tout le test, on dira qu'il y a censure à droite lors du dernier relevé. Finalement, si le sujet apprécie le produit, puis ne l'apprécie plus, puis l'apprécie à nouveau, on considérera qu'il y a une censure à gauche lors du second changement d'avis.

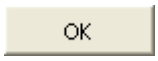
Pour plus de détails sur les modèles paramétriques de survie et sur la censure, on peut voir le chapitre de cette aide dédié à ces méthodes. XLSTAT- MX permet d'utiliser un modèle exponentiel, de Weibull ou log-normal.

En sorties, on trouvera des graphiques ainsi que les estimations des paramètres du modèle.

XLSTAT-MX permet aussi d'ajouter des informations externes en utilisant des variables explicatives associées aux sujets.

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.

Onglet **Général**:

Deux formats de données sont disponibles (voir partie description de cette aide).

Pour le cas « une colonne par date » :

Tableau Sujet x Date : sélectionnez le tableau correspondant à l'avis des sujets pour chacune des dates étudiées. Si un en-tête a été sélectionné sur la première ligne, veillez à ce que l'option « libellés des colonnes » soit activée.

Données de dates : sélectionnez les données correspondant aux dates relevées. Ces données doivent être numériques. Le nombre de ligne doit être égal au nombre de colonne du tableau précédent. Si un en-tête a été sélectionné sur la première ligne, veillez à ce que l'option « libellés des colonnes » soit activée

Code positif : entrez le code utilisé pour identifier qu'un sujet a apprécié le produit à une date donnée. La valeur par défaut est 1.

Code négatif : entrez le code utilisé pour identifier qu'un sujet n'a pas apprécié le produit à une date donnée. La valeur par défaut est 0.

Pour le cas « une ligne par évènement » :

Données de dates : sélectionnez les données correspondant aux dates auxquelles sont notés les changements de situation. Si un en-tête a été sélectionné sur la première ligne, veillez à ce que l'option « libellés des colonnes » soit activée.

Sujets : sélectionnez ici les données correspondant au nom du sujet associé à l'évènement. Si un en-tête a été sélectionné sur la première ligne, veillez à ce que l'option « libellés des colonnes » soit activée

Variables explicatives :

Quantitatives : activez cette option si vous voulez inclure une ou plusieurs variables explicatives quantitatives dans le modèle. Sélectionnez alors la ou les variables correspondantes sur la feuille Excel. Les données sélectionnées doivent être de type numérique. Si le libellé des variables a été sélectionné, veuillez vérifier que l'option « Libellés des colonnes » est activée.

Qualitatives : activez cette option si vous voulez inclure une ou plusieurs variables explicatives qualitatives dans le modèle. Sélectionnez alors la ou les variables correspondantes sur la feuille Excel. Les données sélectionnées peuvent être de tout type, mais les données numériques sont automatiquement considérées comme nominales. Si le libellé des variables a été sélectionné, veuillez vérifier que l'option « Libellés des colonnes » est activée (voir *description*).

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des colonnes : activez cette option si la première ligne des données sélectionnées (variables de temps, de censure et explicatives) contient un libellé.

Distribution : choisissez la distribution que vous voulez appliquer à votre modèle.

Libellés des sujets : dans le cas du format « une colonne par date », activez cette option pour sélectionner les noms des sujets.

Onglet **Options**:

Niveau de signification (%) : entrez la valeur du niveau de signification à utiliser pour les tests (valeur par défaut : 5%). Cette valeur est aussi utilisée pour déterminer les intervalles de confiance pour les statistiques calculées.

Tolérance : activez cette option pour permettre à l'algorithme de calcul de ne pas prendre en compte les variables qui seraient soit constantes soit trop corrélées avec d'autres variables déjà utilisées dans le modèle (0.0001 par défaut).

Valeurs de départ : activez cette option pour donner un point de départ à XLSTAT. Sélectionnez alors les cellules correspondant aux valeurs initiales des paramètres. Le nombre de lignes sélectionnées doit correspondre au nombre de paramètres.

Contraintes : dans le cas de variables explicatives qualitatives, on aura différentes contraintes sur celles-ci :

$a_1 = 0$: choisissez cette option pour que le paramètre de la première modalité de chaque facteur soit fixé à 0.

$a_n = 0$: choisissez cette option pour que le paramètre de la dernière modalité de chaque facteur soit fixé à 0.

Conditions d'arrêt :

- **Itérations** : entrez le nombre maximal d'itérations pour l'algorithme de Newton-Raphson. Les calculs sont interrompus dès que le nombre maximal d'itérations est dépassé. Valeur par défaut : 100.
- **Convergence** : entrez la valeur seuil d'évolution maximale de log vraisemblance d'une itération à l'autre, qui une fois atteinte permet de considérer que l'algorithme a convergé. Valeur par défaut : 0,000001.

Sélection du modèle : activez cette option si vous souhaitez utiliser l'une des deux méthodes de sélection proposées :

- **Ascendante** : le processus de sélection commence par l'ajout de la variable ayant la plus forte contribution au modèle. Si une seconde variable est telle que sa probabilité d'entrée est supérieure à la **valeur seuil pour entrer**, alors elle est ajoutée au modèle.

- **Descendante** : cette méthode est similaire à la précédente, mais part d'un modèle complet.

Onglet **Données manquantes** :

Ne pas accepter les valeurs manquantes : activez cette option pour que XLSTAT empêche la poursuite des calculs si des valeurs manquantes sont détectées.

Supprimer les observations : activez cette option pour supprimer les observations comportant des données manquantes.

Onglet **Sorties**:

Statistiques descriptives : activez cette option pour afficher les statistiques descriptives pour les variables sélectionnées.

Statistiques par date : activez cette option pour afficher les statistiques descriptives pour chaque date étudiée.

Coefficients d'ajustement : activez cette option pour afficher le tableau des statistiques d'ajustement du modèle.

Coefficients du modèle : activez cette option pour afficher le tableau des paramètres du modèle.

Résidus et prédictions : activez cette option pour afficher les différents types de résidus pour l'ensemble des observations (résidus standardisés et résidus de Cox-Snell). Dans ce tableau sont aussi affichées les valeurs prédites pour la fonction de préférence cumulée.

Quantiles : activez cette option pour afficher les quantiles prédits pour l'ensemble de la courbe de survie. XLSTAT-MX donne les quantiles à 1, 5, 10, 25, 50, 75, 90, 95, et 99 %.

Onglet **Graphiques** :

Evolution des préférences : activez cette option pour afficher le graphique relatif à l'évolution des préférences des sujets en fonction du temps.

Fonction de préférence cumulée : activez cette option pour afficher les graphiques relatifs à la fonction de préférence cumulée obtenue avec la distribution sélectionnée.

Résidus : activez cette option si vous souhaitez que XLSTAT affiche le graphique des résidus en fonction du temps.

Résultats

XLSTAT propose un nombre important de tableaux et de graphiques afin de faciliter l'analyse et l'interprétation des résultats.

Sujets retirés de l'analyse : ce tableau donne le nom des sujets retirés de l'analyse car les données en entrée pour ceux-ci n'étaient pas adaptées.

Statistiques descriptives : le tableau de statistiques descriptives présente pour toutes les variables sélectionnées des statistiques simples. Pour les variables quantitatives, sont affichés le nombre d'observations, le nombre de données manquantes, le nombre de données non manquantes, la moyenne, et l'écart-type (non biaisé). Pour les variables qualitatives sont affichées les modalités leurs effectifs et pourcentage respectifs.

Statistique par date : le tableau des statistiques par date donne pour chaque date analysée le nombre de sujets ayant apprécié le produit ainsi que le pourcentage associé.

Synthèse de la sélection des variables : dans le cas où une méthode de sélection a été choisie, XLSTAT affiche la synthèse de la sélection. Dans le cas d'une sélection pas à pas, les statistiques correspondant aux différentes étapes sont affichées.

Coefficients d'ajustement : dans ce tableau est affichée une série de statistiques pour le modèle indépendant (correspondant au cas où il n'y aurait aucune variables dans le modèle) et pour le modèle ajusté.

- **Observations** : le nombre total d'observations prises en compte ;
- **DDL** : degrés de liberté ;
- **-2 Log(Vrais.)** : le logarithme de la fonction de vraisemblance associée au modèle ;
- **AIC** : le critère d'information d'Akaike (Akaike's Information Criterion) ;
- **SBC** : le critère bayésien de Schwarz (Schwarz's Bayesian Criterion) ;
- **Itérations** : nombre d'itérations nécessaires à la convergence de l'algorithme.

Paramètres du modèle : pour chaque paramètre du modèle sont affichés l'estimation du paramètre, l'écart-type correspondant, le Khi^2 de Wald, la p-value correspondante, ainsi qu'un intervalle de confiance associé.

Les **résidus** sont donnés pour chaque observation.

Les **quantiles** sont pour chaque valeur des quantiles pour les courbes de préférence.

Les **graphiques** obtenus dépendent des options activées. On peut représenter la fonction de préférence cumulée ainsi que les graphiques des résidus.

Exemple

Un exemple d'analyse sensorielle de durée de vie est disponible sur le Centre d'aide XLSTAT :

http://www.xlstat.com/demo-shelflifef.htm

Bibliographie

Cox D. R. and Oakes D. (1984). Analysis of Survival Data. Chapman and Hall, London.

Hough G. (2010). Sensory Shelf Life Estimation of Food Products, CRC Press.

Kalbfleisch J. D. and Prentice R. L. (2002). The Statistical Analysis of Failure Time Data. 2^{nd} edition, John Wiley & Sons, New York.

Modèle de Bradley-Terry généralisé

Cet outil permet d'ajuster le modèle de Bradley-Terry sur des données issues de comparaisons par paires.

Dans cette section :

Description

Le modèle de Bradley-Terry généralisé permet de décrire les résultats possibles lorsque des éléments sont mis en comparaison par paires. Par exemple, dans le cadre d'une étude marketing, k produits sont évalués par des consommateurs. Les produits sont soumis par paires et les consommateurs doivent à chaque fois indiquer quel est le produit préféré ou éventuellement dire s'ils ne peuvent se prononcer.

Le modèle de Bradley-Terry généralisé

Soit i et j deux éléments que l'on compare. Bradley et Terry (1952) ont proposé le modèle suivant pour définir la probabilité que i soit supérieur à (ou « batte ») j :

$$P(i > j) = \frac{\lambda_i}{\lambda_i + \lambda_j}$$

où λ_i caractérise l'habileté de l'élément i, $\lambda_i \geq 0$. Plus ce paramètre est grand et plus l'élément i aura de chance d'être supérieur à un élément j.

Plusieurs extensions de ce modèle ont été proposées pour prendre en compte l'avantage à domicile (Agresti (1990), dans le cadre de matchs de sport) ou la possibilité d'avoir une égalité entre deux éléments i et j (Rao et Kupper (1967)).

Lorsque l'on souhaite prendre en compte le fait que l'ordre entre i et j a une importance, Agresti (1990) a proposé d'ajouter un paramètre δ qui mesure la force de l'avantage à domicile. Dans ce cas, le modèle devient :

$$P(i>j)=\begin{cases}\frac{\delta\lambda_i}{\delta\lambda_i+\lambda_j} \text{ si i est à domicile}\\ \\ \frac{\lambda_i}{\lambda_i+\delta\lambda_j} \text{ si j est à domicile}\end{cases}$$

De même, si l'on souhaite autoriser l'égalité entre deux éléments i et j, Rao et Kupper (1967) ont proposé d'incorporer un paramètre $\theta > 1$ dans le modèle usuel tel que :

$$P(i>j)=\frac{\lambda_i}{\lambda_i+\theta\lambda_j}$$

$$P(i=j)=\frac{\left(\theta^2-1\right)\lambda_i\lambda_j}{\left(\lambda_i+\theta\lambda_j\right)\left(\theta\lambda_i+\lambda_j\right)}$$

Inférence des paramètres du modèle

Dans le cas du modèle de Bradley-Terry usuel, un estimateur du maximum de vraisemblance des paramètres peut être obtenu par un simple algorithme itératif MM (Maximisation-Minimisation, Hunter (2004)). On peut également montrer que le modèle avec ou sans avantage à domicile peut être réécrit comme un modèle de régression logistique. Dans ce cas, un algorithme numérique permet de déterminer un estimateur des paramètres.

En 2012, Caron et Doucet ont proposé une approche bayésienne qui, en considérant les paramètres comme des variables aléatoires, permet de contourner des difficultés liées à la faible densité des données. Dans ce contexte, deux approches peuvent être utilisées :

- Maximiser la vraisemblance par un algorithme de type EM. On peut montrer que, pour un choix spécifique de distribution *a priori* sur les paramètres, cet algorithme correspond à l'algorithme originel MM.
- Estimer une distribution *a posteriori* des paramètres par un algorithme d'échantillonnage de type Gibbs.

Ces deux approches s'appuient sur l'introduction de variables latentes telles que la vraisemblance complète s'écrive simplement. Tout d'abord, on définit ω_{ij} comme le nombre de fois où i a battu j, $\omega_i = \sum_{j=1, j\neq i}^{K} \omega_{ij}$ le nombre de victoires de l'élément i et $n_{ij} = \omega_{ij} + \omega_{ji}$ le nombre de comparaisons total entre les éléments i et j. A partir de l'interprétation de Thurstone (Diaconis (1988)), le modèle de Bradley-Terry s'écrit :

$$P(Y_{ki} < Y_{kj}) = \frac{\lambda_i}{\lambda_i+\lambda_j}$$

avec $Y_{ki} \sim \epsilon(\lambda_i)$ et $k \in \{1, \ldots, n_{ij}\}$. Afin de simplifier la vraisemblance complète du modèle, on introduit une nouvelle variable latente Z_{ij} définie telle que :

$$Z_{ij} = \sum_{k=1}^{n_{ij}} min(Y_{kj}, Y_{ki}) \sim \Gamma(n_{ij}, \lambda_i + \lambda_j)$$

Dans un contexte bayésien, une distribution *a priori* est associée à chacun des paramètres. On suppose alors que les paramètres λ_i sont distribués selon une loi Gamma de paramètres a et b :

$$P(\lambda) = \prod_{i=1}^{K} \Gamma(\lambda_i; a, b)$$

De même, le paramètre d'avantage à domicile δ est distribué selon une loi $\Gamma(\delta; a_\delta, b_\delta)$ et le paramètre d'égalité θ selon une distribution *a priori* impropre sur $[1, +\infty[$.

EM Bayésien : cette approche itérative vise à maximiser l'espérance de la log-vraisemblance conditionnellement aux données.

Modèle usuel :

L'estimateur du paramètre λ_i à l'itération t s'écrit :

$$\lambda_i^{(t)} = \frac{a - 1 + \omega_i}{b + \sum_{j \neq i} \frac{n_{ij}}{\lambda_i^{(t-1)} + \lambda_j^{(t-1)}}}$$

Si $a = 1$ et $b = 0$ alors on retrouve le même estimateur que dans l'algorithme MM.

Modèle avec avantage à domicile :

Les estimateurs des paramètres λ_i et δ à l'itération t s'écrivent :

$$\lambda_i^{(t)} = \frac{a - 1 + \omega_i}{b + \sum_{j \neq i} \frac{\delta^{(t-1)} n_{ij}}{\delta^{(t-1)} \lambda_i^{(t-1)} + \lambda_j^{(t-1)}} + \frac{n_{ji}}{\lambda_i^{(t-1)} + \delta^{(t-1)} \lambda_j^{(t-1)}}}$$

et

$$\delta^{(t)} = \frac{a_\delta - 1 + c}{b_\delta + \sum_{j \neq i} \frac{\lambda_i^{(t)} n_{ij}}{\delta^{(t-1)} \lambda_i^{(t-1)} + \lambda_j^{(t-1)}}}$$

Avec $c = \sum_{i \neq j} a_{ij}$ où a_{ij} représente le nombre de victoires de i sur j quand i est à domicile.

Modèle avec égalités :

On définit par t_{ij} le nombre d'égalités entre les éléments i et j. Les estimateurs des paramètres λ_i et θ à l'itération t s'écrivent :

$$\lambda_i^{(t)} = \frac{a - 1 + s_i}{b + \sum_{j \neq i} \frac{s_{ij}}{\lambda_i^{(t-1)} + \Theta^{(t-1)} \lambda_j^{(t-1)}} + \frac{\Theta^{(t-1)} s_{ji}}{\Theta^{(t-1)} \lambda_i^{(t-1)} + \lambda_j^{(t-1)}}}$$

Où $s_{ij} = \omega_{ij} + t_{ij}$ et $s_i = \sum_{j \neq i} s_{ij}$. On a :

$$\Theta^{(t)} = \frac{1}{2c^{(t)}} + \sqrt{1 + \frac{1}{4c^{(t)^2}}}$$

et

$$c^{(t)} = \frac{2}{T} \sum_{j \neq i} \frac{s_{ij} \lambda_j^{(t)}}{\lambda_i^{(t-1)} + \Theta^{(t-1)} \lambda_j^{(t-1)}}$$

avec $T = \frac{1}{2} \sum_{j \neq i} t_{ij}$ le nombre total d'égalités.

Echantillonnage : cette approche repose sur l'utilisation de l'algorithme de Gibbs.

Modèle usuel :

L'algorithme utilisé pour obtenir un estimateur de λ_i est le suivant :

Pour $1 \leq i < j \leq K$ s.t. $n_{ij} > 0$,

$$Z_{ij}^{(t)} | X, \lambda^{(t-1)} \sim \Gamma \left(n_{ij}, \lambda_i^{(t-1)} + \lambda_j^{(t-1)} \right)$$

Pour $1 \leq i \leq K$,

$$\lambda^{(t)} | X, Z^{(t)} \sim \Gamma(a + \omega_i, b + \sum_{i<j | n_{ij}>0} Z_{ij}^{(t)} + \sum_{j<i | n_{ij}>0} Z_{ji}^{(t)})$$

Modèle avec avantage à domicile :

L'algorithme utilisé pour obtenir un estimateur de λ_i et de δ est le suivant : Pour $1 \leq i < j \leq K$ s.t. $n_{ij} > 0$,

$$Z_{ij}^{(t)} | X, \lambda^{(t-1)}, \delta^{(t-1)} \sim \Gamma \left(n_{ij}, \delta^{(t-1)} \lambda_i^{(t-1)} + \lambda_j^{(t-1)} \right)$$

Pour $1 \leq i \leq K$,

$$\lambda^{(t)}|X, Z^{(t)}, \delta^{(t-1)} \sim \Gamma(a + \omega_i, b + \delta^{(t-1)} \sum_{i \neq j | n_{ij} > 0} Z_{ij}^{(t)} + \sum_{j \neq i | n_{ij} > 0} Z_{ji}^{(t)})$$

Puis,

$$\delta^{(t)}|X, Z^{(t)}, \lambda^{(t-1)} \sim \Gamma \left(a_\delta + c, b_\delta + \sum_{i=1}^{K} \lambda_i^{(t)} \sum_{j \neq i | n_{ij} > 0} Z_{ij}^{(t)} \right)$$

Modèle avec égalités :

L'algorithme utilisé pour obtenir un estimateur de λ_i et de θ est le suivant :

Pour $1 \leq i < j \leq K$ s.t. $n_{ij} > 0$,

$$Z_{ij}^{(t)}|X, \lambda^{(t-1)}, \Theta^{(t-1)} \sim \Gamma \left(s_{ij}, \lambda_i^{(t-1)} + \Theta^{(t-1)} \lambda_j^{(t-1)} \right)$$

Pour $1 \leq i \leq K$,

$$\lambda^{(t)}|X, Z^{(t)}, \Theta^{(t-1)} \sim \Gamma(a + s_i, b + \sum_{i \neq j | s_{ij} > 0} Z_{ij}^{(t)} + \Theta^{(t-1)} \sum_{j \neq i | n_{ij} > 0} Z_{ji}^{(t)})$$

Puis,

$$\Theta^{(t)}|X, Z^{(t)}, \lambda^{(t)} \sim P \left(\Theta | X, Z^{(t)}, \lambda^{(t)} \right)$$

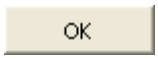
avec :

$$P \left(\Theta | X, Z^{(t)}, \lambda^{(t)} \right) \propto \left(\Theta^2 - 1 \right)^T exp \left(- \sum_{i \neq j | s_{ij} > 0} Z_{ij} \Theta \right)$$

Ces deux approches mènent à la construction d'une distribution *a posteriori* sur les paramètres du modèle. Cependant, uniquement la seconde (échantillonnage de Gibbs) permet d'inférer les paramètres du modèle complet (avantage à domicile et égalité).

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

Annuler : cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

Aide : cliquez sur ce bouton pour afficher l'aide.

 : cliquez sur ce bouton pour rétablir les options par défaut.

 : cliquez sur ce bouton pour effacer les sélections de données.

 : cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.

 : cliquez sur ce bouton pour changer la façon dont vous chargez les données dans XLSTAT. Si le bouton possède une icône de souris, XLSTAT vous permet de faire une sélection à la souris des données. Si le bouton possède une icône de liste, XLSTAT vous permet de sélectionner les données à partir d'une liste. Si le bouton possède une icône orange, des boutons complémentaires avec un point d'interrogation  sont affichés pour vous permettre d'importer les données à partir de fichiers.

Onglet **Général** :

Format des données : choisissez le format des données.

- **Tableau croisé** : activez cette option pour sélectionner des données présentées sous la forme d'un tableau de contingence où les victoires sont en lignes et les défaites en colonne. Dans ce cas, uniquement le modèle classique est envisageable.
- **Tableau Paires/Variables** : activez cette option pour sélectionner des données présentées sous la forme de deux tableaux. Le tableau des paires correspond à l'ensemble des rencontres entre les différents éléments. Le tableau des variables correspond aux résultats des rencontres. La première colonne représente les victoires du premier élément et la seconde ses défaites. Une troisième colonne optionnelle peut contenir le nombre d'égalités entre les deux éléments.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Etiquettes : activez cette option si des en-têtes de colonne ont été sélectionnés (ou des en-têtes de ligne en mode lignes).

Onglet **Options** :

Méthode d'inférence : sélectionnez la méthode avec laquelle l'inférence des paramètres sera réalisée (voir la section description).

Numérique : le modèle est réécrit sous la forme d'une régression logistique. L'égalité entre les éléments n'est pas autorisée pour cette méthode d'inférence.

EM Bayésien : les paramètres sont supposés être distribués *a priori* selon une distribution Gamma. L'inférence est réalisée par un algorithme EM permettant la mise à jour des différentes distributions. Avec cette méthode, l'inférence du modèle complet ne peut pas être réalisée.

Echantillonnage : les paramètres sont supposés être distribués *a priori* selon une distribution Gamma. La distribution *a posteriori* de chaque paramètre est obtenue via un algorithme de Gibbs. L'ensemble des modèles peut être envisagé avec cette méthode d'inférence.

Options de modèle :

Avantage à domicile : sélectionnez cette option pour que l'avantage à domicile soit pris en compte dans le modèle. Dans ce cas, l'ordre des rencontres contenues dans le tableau des paires a une importance. Le premier élément est supposé être à domicile.

Egalité : sélectionnez cette option si le résultat de la comparaison entre deux éléments peut être l'égalité. Si cette option est activée, le tableau des variables doit être constitué de 3 colonnes. Attention, seule la méthode d'échantillonnage permet de prendre en compte à la fois l'avantage à domicile et l'égalité.

Intervalle de confiance (%) : entrez l'étendue en pourcentage de l'intervalle de confiance à fournir pour les paramètres.

Conditions d'arrêt :

Itérations /Nombre de simulations : nombre maximal d'itérations autorisé pour l'algorithme d'inférence.

Temps maximum : temps maximum autorisé pour l'inférence des paramètres (en secondes).

Convergence : seuil de convergence.

Paramètres *a priori* : cette option est active uniquement si la méthode d'inférence choisie est EM Bayésien ou échantillonnage.

Echelle : paramètre d'échelle de la distribution Gamma.

Forme : paramètre de forme de la distribution Gamma.

Onglet **Sorties** :

Statistiques descriptives : activez cette option pour calculer et afficher les statistiques descriptives pour les différentes variables.

Critère de vraisemblance : activez cette option pour calculer et afficher la vraisemblance, les critères BIC (Bayesian Information Criterion) et AIC (Akaike Information Criterion).

Probabilités de victoire : activez cette option pour calculer et afficher les probabilités de victoire d'un élément sur l'autre, en fonction des options choisies.

Formules Excel : activez cette option pour afficher la formule de calcul de probabilité lorsque vous cliquez sur une case du tableau de probabilités.

Onglet **Graphiques** :

Graphiques de convergence : activez cette option afficher les graphiques de convergence lorsque l'inférence est réalisée par la méthode d'échantillonnage.

Balloon plot : activez cette option pour afficher un graphique permettant d'analyser rapidement les probabilités.

Résultats

Statistiques descriptives : dans ce tableau sont affichées les statistiques descriptives correspondant aux différents éléments.

Paramètres estimés : les paramètres estimés du modèle sont présentés dans un tableau. L'écart-type et un intervalle de confiance sont également fournis pour chacun des paramètres.

Critères de vraisemblance : dans ce tableau, plusieurs critères de vraisemblance sont présentés (-2*log(Vraisemblance), BIC, AIC).

Probabilités de victoire : à partir des paramètres du modèle, un tableau présentant la probabilité qu'un élément i (en ligne) batte l'élément j (en colonne) est fourni. Le cas échéant, la probabilité d'une égalité ainsi qu'une différenciation entre domicile et extérieur sont également présentées.

Graphiques de convergence : pour chacun des paramètres du modèle, l'évolution de la valeur du paramètre au cours des itérations de l'algorithme avec un intervalle de confiance est représenté.

Balloon plot : pour chacun des tableaux de probabilités, un graphique montrant les disparités par la taille des cercles et les valeurs par les couleurs est affiché.

Exemple

Un exemple d'application du modèle de Bradley-Terry est disponible sur le Centre d'aide XLSTAT :

http://www.xlstat.com/demo-bradleyf.htm

Bibliographie

Bradley R. and Terry M. (1952). Rank analysis of incomplete block designs. I. the method of paired comparisons. *Biometrika*, **39**, 324-345.

Caron F. and Doucet A. (2012). An Efficient Bayesian inference or generalized Bradley-Terry models. *Journal of Computational and Graphical Statistics*, to be published.

Diaconis P. (1988). Group representations in probability and statistics, *IMS Lecture Notes*, **11**. Institute of Mathematical Statistics.

Hunter D. (2004). MM algorithms for generalized Bradley-Terry models. *The Annals of Statistics*, **32**, 384-406.

Rao P. and Kupper L. (1967). Ties in paired-comparison experiments: A generalization of the Bradley-Terry model. *Journal of the American Statistical Association*, **62**, 194–204.

Analyse Procrustéenne Généralisée

Utilisez l'analyse procrustéenne généralisée (*Generalized Procrustes Analysis* ou GPA en anglais) pour transformer plusieurs configurations multidimensionnelles de manière à les rendre le plus semblables possible et pour éventuellement ensuite comparer les configurations transformées.

Dans cette section :

Description

Procruste (ou Procuste), qui en grec ancien signifie « celui qui allonge en tirant », est un personnage de la mythologie grec. Le nom du bandit Procruste est associé au lit de torture dont il se servait pour supplicier les voyageurs auxquels il proposait le gîte. Procruste installait sa future victime sur un lit à dimensions variables : court pour les grands et long pour les petits. Selon le cas, il tranchait d'un coup d'épée ce qui dépassait du lit ou allongeait le corps du voyageur jusqu'à amener la longueur du malheureux à celle du lit, en utilisant un mécanisme qu'Héphaïstos lui avait fabriqué. Thésée anticipa le piège et se mit dans le lit en biais. Lorsque Procuste vint ajuster le corps de Thésée, il ne comprit pas immédiatement la situation et resta perplexe le temps pour Thésée de sectionner, d'un coup d'épée, le brigand en deux parties égales.

L'analyse procrustéenne généralisée (*Generalised Procrustes Analysis* ou GPA en anglais) est une méthode mathématique qui permet de réaliser des transformations sur des tableaux multidimensionnels de manière à réduire la distance euclidienne entre ces tableaux.

L'analyse procrustéenne généralisée est souvent utilisée en analyse sensorielle en préalable à une cartographie des préférences (Preference mapping) par exemple pour réduire les effets d'échelles et pour aboutir à une configuration consensuelle. Elle peut aussi permettre d'analyser la proximité de certains termes utilisés par différents experts.

Principe

On désigne par configuration une matrice n x p (n objets, p dimensions) correspondant à la description de n objets (ou individus/produits) suivant p dimensions (ou attributs/ variables/critères/descripteurs).

On appelle configuration consensuelle la configuration moyenne calculée à partir des m configurations. L'analyse procrustéenne généralisée est une méthode itérative qui permet de réduire par une suite de transformations des m configurations (changement d'échelle, translations, rotations, réflexions), la distance des m configurations à la configuration consensuelle, cette dernière évoluant après chaque transformation.

Prenons l'exemple de 5 experts notant 4 fromages suivant 3 critères, les notes pouvant aller de 1 à 10. On peut facilement envisager qu'un sujet ait tendance à être plus dur dans sa notation, entraînant un décalage vers le bas des notes, ou qu'un autre ait tendance à mettre des notes autour de la moyenne, sans oser se risquer à utiliser des notes extrêmes. Travailler sur une configuration moyenne risquerait alors d'entraîner de fausses interprétations. On comprend aisément qu'une translation des notes du premier sujet est nécessaire, ou qu'une remise à l'échelle des notes du second sujet rendrait les notes de ce dernier éventuellement plus proches de celles des autres sujets.

Une fois la configuration consensuelle obtenue, il est possible de réaliser une ACP de manière à permettre une visualisation optimale en deux ou trois dimensions des configurations après transformation et de la configuration consensuelle. XLSTAT-MX réalise une ACP non normée et affiche le cercle des corrélations et la carte des objets.

Structure des données

Il existe deux cas différents :

1. Si le nombre et la désignation des p dimensions sont identiques pour les m configurations, on parle en analyse sensorielle de profils conventionnels.

2. Si le nombre p et la désignation des dimensions varie d'une configuration à l'autre, on parle en analyse sensorielle de profils libres, et les données ne peuvent alors être représentées que sous la forme d'une suite de m matrices de taille n x p(k), k=1,2, …, m.

Pour la saisie des données, XLSTAT vous demande de sélectionner un tableau n x (p x m), correspondant aux m configurations contiguës. Si les données ne sont **pas adaptées** à ce type de tableau, il est possible de les **transformer** en un tableau horizontal en utilisant la fonctionnalité **Créer un tableau Produits\Sujets**.

Transposition des données

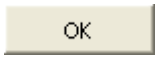
Il arrive fréquemment que le nombre (m x p) de colonnes du tableau des configurations dépasse la limite imposée par Excel. Pour pallier ce problème, XLSTAT vous permet d'utiliser des tableaux transposés. Pour utiliser des tableaux transposés (tous les tableaux sélectionnés doivent alors être transposés), il vous suffit que cliquer sur le bouton de transposition : la flèche bleue en bas à gauche de la boîte de dialogue devient alors rouge.

Algorithmes

XLSTAT est le seul logiciel offrant le choix entre les deux principaux algorithmes disponibles, le premier fondé sur les travaux initiés par John Gower (1975), et le second basé sur les travaux de Jacques Commandeur (1991). En fonction du jeu de données, l'un ou l'autre algorithme sera le plus performant (en termes de moindres carrés), mais l'algorithme de Commandeur a la particularité de permettre de prendre en compte des données manquantes. Par données manquantes, on entend ici que pour une configuration donnée et une observation donnée, les valeurs n'ont pas été enregistrées pour toutes les dimensions de la configuration. Ce dernier cas peut se produire en analyse sensorielle, si l'un des sujets n'a pas évalué d'un produit.

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.

Onglet **Général** :

Configurations : sélectionnez les données correspondant aux configurations. Si la première ligne de la sélection comprend des en-têtes, l'option « Libellés des dimensions » doit être activée. Si les données ne sont **pas adaptées** à ce type de tableau, il est possible de les **transformer** en un tableau horizontal en utilisant la fonctionnalité **Créer un tableau Produits\Sujets**.

Nombre de configurations : entrez le nombre de configurations contenues dans la sélection ci-dessus.

Nombre de dimensions par configuration :

- **Egal** : choisissez cette option si le nombre de dimensions est identique pour toutes les configurations. XLSTAT détermine alors automatiquement le nombre de dimensions de

chacune des configurations.

- **Défini par l'utilisateur** : choisissez cette option pour sélectionner une plage contenant les nombres de dimensions correspondant à chacune des configurations. Si l'option « Libellés des dimensions » est activée, la première cellule de la sélection doit comprendre un en-tête.

Libellés des configurations : activez cette option si vous voulez utiliser les libellés des configurations pour l'affichage des résultats. Si l'option « Libellés des dimensions » est activée, la première cellule de la sélection doit comprendre un en-tête. Si vous n'activez pas cette option, des libellés seront automatiquement créés (C1, C2, …).

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des dimensions : activez cette option si la première ligne des données sélectionnées (configurations, libellés des configurations, libellés des objets) contient un libellé.

Libellés des objets : activez cette option si vous voulez utiliser des libellés d'objets pour l'affichage des résultats. Si l'option « Libellés des dimensions » est activée, la première cellule de la sélection doit comprendre un en-tête. Si vous n'activez pas cette option, des libellés seront automatiquement créés (Obs1, Obs2, …).

Méthode : choisissez quel algorithme doit être utilisé :

- **Commandeur** : activez cette option pour utiliser l'algorithme de Commandeur (voir la section description pour plus de détails).
- **Gower** : activez cette option pour utiliser l'algorithme de Gower (voir la section description pour plus de détails).

Onglet **Options** :

Mise à l'échelle : activez cette option pour effectuer les mises à l'échelle (*rescaling*).

Rotation/Réflexion : activez cette option pour effectuer les rotations/réflexion.

ACP : activez cette option pour effectuer une Analyse en Composantes Principales en fin d'analyse.

Filtrer les facteurs : vous pouvez activer l'une ou les deux options suivantes afin de réduire le nombre de facteurs pris en compte à la suite de l'ACP :

- **% minimum** : activez cette option puis saisissez le pourcentage minimum de la variabilité totale que doivent représenter les facteurs retenus.
- **Nombre maximum** : activez cette option pour fixer le nombre maximum de facteurs à prendre en compte.

Tests :

- **Test de consensus** : activez cette option pour utiliser un test de permutation permettant de déterminer si un consensus est obtenu à la suite des transformations.
- **Test de dimensions** : activez cette option pour utiliser un test de permutation permettant de déterminer quel est le bon nombre de facteurs à retenir.

Nombre de permutations : entrez le nombre de permutations à réaliser pour les tests (valeur par défaut : 300)

Niveau de signification (%) : entrez le niveau de signification pour les tests.

Conditions d'arrêt :

- **Itérations** : entrez le nombre maximal d'itérations pour l'algorithme. Les calculs sont interrompus dès que le nombre maximal d'itérations est dépassé. Valeur par défaut : 100.
- **Convergence** : entrez la valeur seuil d'évolution maximale du critère de convergence d'une itération à l'autre, qui une fois atteinte permet de considérer que l'algorithme a convergé. Valeur par défaut : 0,00001.

Onglet **Données manquantes** :

Ne pas accepter les valeurs manquantes : activez cette option pour que XLSTAT empêche la poursuite des calculs si des valeurs manquantes sont détectées.

Remplacer les données manquantes : activez cette option pour remplacer les données manquantes par 0.

Supprimer les observations : activez cette option pour supprimer les observations comportant des données manquantes.

Ignorer les données manquantes : activez cette option pour ignorer les données manquantes. La méthode utilisée amène à considérer que si une valeur est manquante pour l'objet J et pour la dimension D de la configuration C, alors les valeurs sont manquantes pour l'objet J pour toutes les dimensions de la configuration C. En revanche, les valeurs pour l'objet J pour les autres configurations ne sont pas affectées.

Onglet **Sorties** :

Tableau de PANOVA : activez cette option pour afficher le tableau de PANOVA.

Résidus par objet : activez cette option pour afficher les résidus pour chacun des objets.

Résidus par configuration : activez cette option pour afficher les résidus pour chacune des configurations.

Facteurs de mise à l'échelle : activez cette option pour afficher les facteurs de mises à l'échelle appliqués à chacune des configurations.

Matrices de rotation : activez cette option pour afficher les matrices de rotation associées à chaque configuration.

Les options ci-dessous ne sont disponibles que si une ACP a été demandée :

Valeurs propres : activez cette option pour afficher les valeurs propres de l'ACP.

Configuration consensus : activez cette option pour afficher les coordonnées des dimensions pour la configuration consensus (ou configuration moyenne).

Configurations : activez cette option pour afficher les coordonnées des dimensions pour chacune des configurations.

Coordonnées des objets : activez cette option pour afficher les coordonnées objets après les transformations.

- **Présentation par configuration** : activez cette option pour afficher un tableau de coordonnées par configuration.
- **Présentation par objet** : activez cette option pour afficher un tableau de coordonnées par objet.

Onglet **Graphiques (ACP)** :

Les options ci-dessous ne sont disponibles que si une ACP a été demandée :

Valeurs propres : activez cette option pour afficher le diagramme en bâtons des valeurs propres de l'ACP.

Graphiques de corrélations : activez cette option pour afficher les cercles des corrélations pour la configuration consensus et pour les configurations individuelles.

- **Vecteurs** : activez cette option pour utiliser des vecteurs.

Coordonnées des objets : activez cette option pour représenter graphiquement les objets.

- **Présentation par configuration** : activez cette option pour afficher un graphique où la couleur dépend de la configuration.
- **Présentation par objet** : activez cette option pour afficher un graphique où la couleur dépend de l'objet.

Full biplot : activez cette option pour afficher le graphique présentant à la fois les objets et les dimensions des différentes configurations.

Etiquettes colorées : activez cette option pour que les étiquettes soient de la même couleur que les points correspondants.

Biplots : activez cette option pour afficher les graphiques de représentation simultanée des observations et des variables.

Type de biplots : choisissez le type de biplot que vous souhaitez afficher. Voir la section description de l'ACP pour plus de détails.

- **Biplot de corrélation** : activez cette option pour afficher des biplots de corrélation.
- **Biplot de distance** : activez cette option pour afficher des biplots de distance.
- **Biplot symétrique** : activez cette option pour afficher des biplots symétriques.
- **Coefficient** : choisissez le coefficient dont la racine carrée sera multipliée par les coordonnées des variables. Ce coefficient vous permettra d'ajuster la position des points variables dans le biplot afin de rendre ce dernier plus lisible. Si ce coefficient est différent de 1, la longueur des vecteurs variables n'est plus interprétable en termes d'écart-type (biplot de corrélation) ou de contribution (biplot de distance).

Onglet **Graphiques** :

Résidus par objet : activez cette option pour afficher le diagramme en bâtons des résidus pour chacun des objets.

Résidus par configuration : activez cette option pour afficher le diagramme en bâtons des résidus pour chacune des configurations.

Facteurs de mise à l'échelle : activez cette option pour afficher le diagramme en bâtons des facteurs de mises à l'échelle appliqués à chacune des configurations.

Histogrammes des tests : activez cette option pour afficher les histogrammes à partir des résultats des tests de permutation.

Résultats

Tableau de PANOVA : inspiré du format du tableau d'analyse de la variable du modèle linéaire, ce tableau permet d'évaluer l'apport respectif des différentes transformations. Dans ce tableau sont présentées la variance résiduelle finale, la variation de variance due à la mise à l'échelle des configurations à la rotation et à la translation. Le calcul de la statistique F de Fisher permet de comparer les contributions relatives des différentes transformations. Les probabilités correspondantes permettent d'évaluer si les transformations ont un effet significatif ou non en termes de réduction de la variance.

Résidus par objet : ce tableau et le diagramme en bâtons correspondant permettent de visualiser la répartition de la variance résiduelle par objet. On peut ainsi repérer pour quels objets la GPA a été moins efficace, autrement dit, quels objets se démarquent le plus de la configuration consensuelle.

Résidus par configuration : ce tableau et le diagramme en bâtons correspondant permettent de visualiser la répartition de la variance résiduelle par configuration. On peut ainsi repérer pour quelles configurations la GPA a été moins efficace, autrement dit, quelles configurations se démarquent le plus de la configuration consensuelle.

Facteurs de mise à l'échelle pour chaque configuration : ce tableau et le diagramme correspondant permettent de comparer les facteurs de mise à l'échelle pour les différentes configurations. Il est utilisé en analyse sensorielle pour comprendre comment les sujets ou experts utilisent différemment les échelles de notation.

Matrices de rotation : les matrices de rotation appliquées à chaque configuration sont affichées si l'utilisateur l'a demandé.

Résultats du test de consensus : dans ce tableau sont affichés, le nombre de permutations effectuées, la valeur Rc qui correspond à la proportion de variance totale expliquée par le consensus, et le quantile correspondant à Rc étant donnée la distribution de Rc obtenue à la suite des permutations. Pour évaluer si la GPA est efficace, on se fixe un intervalle de confiance (typiquement 95%), et si le quantile est au-delà de l'intervalle de confiance, on conclut que la GPA a significativement réduit la variance.

Résultats du test de dimensions : dans ce tableau sont affichés, pour chaque facteur retenu à l'issue de l'ACP, le nombre de permutations effectuées, le F calculé à la suite de la GPA (F est ici le rapport de la variance entre les objets sur la variance entre les configurations), le quantile correspondant au F étant donnée la distribution de F obtenue à la suite des permutations. Pour évaluer si un facteur contribue significativement à la qualité de la GPA, on se fixe un intervalle de confiance (typiquement 95%), et si le quantile est au-delà de l'intervalle de confiance, on conclut que le facteur contribue significativement. A titre indicatif sont aussi affichées les valeurs critiques et les p-values de la distribution F de Fisher pour le niveau alpha choisi. Il se peut que les conclusions issues de la distribution F de Fisher soit très différentes de ce qu'indique le test de permutation : l'utilisation de la distribution F de Fisher suppose la normalité des données, ce qui n'est pas nécessairement le cas.

Résultats pour la configuration consensus :

Coordonnées des objets avant l'ACP : ce tableau correspond aux coordonnées moyennes des objets, après les transformations de la GPA, et avant l'ACP.

Valeurs propres : si une ACP a été demandée, le tableau des valeurs propres et le diagramme en bâtons correspondant sont affichés. De ces valeurs propres est déduit le pourcentage de variabilité totale correspondant à chaque axe.

Corrélations des variables avec les facteurs : ces résultats correspondent aux corrélations entre les variables de la configuration consensus avant les transformations, avec les facteurs

obtenus après les transformations (GPA et ACP si cette dernière a été demandée).

Coordonnées des objets : ce tableau correspond aux coordonnées moyennes des objets, après les transformations de la GPA puis de l'ACP si cette dernière a été demandée. Ces résultats sont utilisés pour la construction du graphique des objets.

Résultats pour les configurations après transformations :

Variance par configuration et par facteur : ce tableau, et le diagramme en bâtons qui lui correspond, permettent de visualiser comment se répartit pour chaque configuration la variance pour chacun des facteurs générés par l'ACP.

Corrélations entre les variables et les facteurs : ces résultats correspondent aux corrélations entre les coordonnées des configurations avant et après les transformations (GPA et ACP si cette dernière a été demandée). Ces résultats sont utilisés pour construire le cercle des corrélations si une ACP a été effectuée. Sur le cercle des corrélations, les libellés explicites des variables utilisées pour chaque configuration sont affichés.

Coordonnées des objets (présentation par configuration) : cette série de tableau correspond aux coordonnées des objets pour chaque configuration, après les transformations de la GPA puis de l'ACP si cette dernière a été demandée. Ces résultats sont utilisés pour la construction de la première série de graphiques des objets.

Coordonnées des objets (présentation par objet) : cette série de tableaux correspond aux coordonnées des objets pour chaque configuration, après les transformations de la GPA puis de l'ACP si cette dernière a été demandée. Ces résultats sont utilisés pour la construction de la seconde série de graphiques des objets.

Exemple

Un exemple d'Analyse Procrustéenne Généralisée est disponible sur le Centre d'aide XLSTAT :

http://www.xlstat.com/demo-gpaf.htm

Bibliographie

Commandeur J.J.F. (1991). Matching Configurations. DSWO Press, Leiden.

Dijksterhuis G.B. and Gower J.C. (1991). The interpretation of generalized procrustes analysis and allied methods. Food Quality and Preference. **3**, 67-87.

Gower J.C. (1975). Generalised Procrustes Analysis. *Psychometrika*, **40** (1), 33-51.

Naes T. and Risvik E. (1996). Multivariate Analysis of Data in Sensory Science. Elsevier Science, Amsterdam.

Rodrigue N. (1999). A comparison of the performance of generalized procrustes analysis and the intraclass coefficient of correlation to estimate interrater reliability. Department of Epidemiology and Biostatistics. McGill University.

Ten Berge J.M.F., Kiers H.A.L. and Commandeur J.J.F. (1993). Orthogonal procrustes rotations for matrices with missing values. *British J. of mathematical and statistical psychology*. **46**, 119-134.

Wakeling I.N., Raats M.M. and MacFie H.J.H. (1992). A new significance test for consensus in generalized Procrustes analysis. *Journal of Sensory Studies*, **7**, 91-96.

Wu W., Gyo Q., de Jong S. and Massart D.L. (2002). Randomisation test for the number of dimensions of the group average space in generalised Procrustes analysis. *Food Quality and Preference*, **13**, 191-200.

Analyse Factorielle Multiple (AFM)

Utilisez l'Analyse Factorielle Multiple (AFM) pour analyser simultanément plusieurs tableaux de variables, et obtenir des résultats, notamment des représentations graphiques, qui permettent d'étudier la relation entre les observations, les variables et les tableaux. Chaque tableau doit contenir un type unique de variable : tableau quantitatif, tableau qualitatif ou tableau de fréquences (données de comptage). Cependant on peut réaliser l'analyse sur plusieurs tableaux de type différents.

Dans cette section :

Description

L'Analyse Factorielle Multiple (AFM), permet d'analyser simultanément plusieurs tableaux de variables, et d'obtenir des résultats, notamment des représentations graphiques, qui permettent d'étudier la relation entre les observations, les variables et les tableaux (Escofier et Pagès, 1984). A l'intérieur d'un tableau les variables doivent être de même nature (quantitative, qualitative ou fréquence), mais les tableaux peuvent être de différents types.

L'AFM est une synthèse de l'ACP (Analyse en Composantes Principales) pour les variables quantitatives, l'ACM (Analyse des Correspondances Multiples) pour les variables qualitatives et de l'AFC (Analyse Factorielle des Correspondances) pour un tableau de fréquences. La méthodologie de l'AFM se décompose en deux étapes :

On réalise successivement pour chacun des tableaux une ACP, une ACM ou une AFC en fonction de la nature des variables. On conserve la valeur de la première valeur propre de chacune des analyses pour pondérer ensuite les différents tableaux dans la seconde partie de l'analyse.

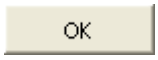
On réalise ensuite une ACP pondérée sur les colonnes de l'ensemble des tableaux, sachant que les tableaux de variables qualitatives sont transformés en tableaux disjonctifs complets, chacune des indicatrices des tableaux disjonctifs ayant un poids fonction de la fréquence de la modalité concernée. La pondération des tableaux permet d'éviter que les tableaux comprenant plus de variables ne pèsent trop dans l'analyse.

Cette méthode s'avère très utile pour analyser des enquêtes lorsque les questions peuvent être regroupées par thèmes, ou lorsque les mêmes questions sont posées à plusieurs intervalles de temps.

Les auteurs ayant développé la méthode (Escofier et Pagès, 1984) ont particulièrement insisté sur l'utilisation des résultats qui découlent de l'AFM. L'originalité première de cette méthode vient du fait qu'elle permet une visualisation dans un espace à deux ou trois dimensions, des tableaux (chaque tableau étant représenté par un point), des variables (dans un cercle des corrélations), des facteurs principaux des analyses de la première phase, et des individus. Par ailleurs, on peut étudier l'impact des autres tableaux sur une observation en visualisant simultanément l'observation décrite par l'ensemble des variables, et par seulement chacun des tableaux. On parle alors de nuages partiels.

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.

Onglet **Général** :

Tableau observations/variables : sélectionnez un tableau comprenant N observations décrites par P variables regroupées dans K tableaux. Si des en-têtes de colonnes ont été sélectionnés, veuillez vérifier que l'option « Libellés des variables » est activée. Si les données ne sont **pas adaptées** à ce type de tableau, il est possible de les **transformer** en un tableau horizontal en utilisant la fonctionnalité **Créer un tableau Produits\Sujets**.

Nombre de tableaux : entrez le nombre K de tableaux constituant le tableau principal des observations variables.

Libellés des tableaux : activez cette option si vous voulez utiliser des libellés pour les K tableaux. Si l'option « Libellés des variables » est activée, la première cellule de la sélection doit comprendre un en-tête. Si vous n'activez pas cette option, des libellés seront automatiquement créés (Tableau1, Tableau2, …).

Nombre de variables par tableau :

- **Egal** : choisissez cette option si le nombre de variables est identique pour tous les tableaux. XLSTAT détermine alors automatiquement le nombre de variables de chacun des tableaux.
- **Défini par l'utilisateur** : choisissez cette option pour sélectionner un vecteur colonne contenant le nombre de variables contenu dans chaque tableau. Si l'option « Libellés des variables » est activée, la première cellule de la sélection doit comprendre un en-tête.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des variables : activez cette option si la première ligne des données sélectionnées (Tableau observations/variables, libellés des observations, poids, …) contient un libellé.

Libellés des observations : activez cette option si vous voulez utiliser des libellés d'observations pour l'affichage des résultats. Si l'option « Libellés des variables » est activée, la première cellule de la sélection doit comprendre un en-tête. Si vous n'activez pas cette option, des libellés seront automatiquement créés (Obs1, Obs2, …).

Poids : activez cette option si vous voulez pondérer les observations. Si vous n'activez pas cette option, les poids seront tous considérés comme valant 1. Les poids doivent être impérativement supérieurs ou égaux à 0. Si un en- tête de colonne a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée. Attention si votre tableau observations/variables contient des données de fréquence, les poids seront automatiquement égaux à 1 même si vous sélectionnez un vecteur de poids.

Onglet **Options** :

Type de données : précisez quel est le type des données des différents tableaux, sachant que le type de données doit être identique à l'intérieur d'un sous-tableau. Dans le cas où le type est mixte, pour indiquer à XLSTAT quel est le type des tableaux, vous devez alors sélectionner une plage indiquant le type des K tableaux. Utilisez 0 pour un tableau contenant des variables quantitatives, 1 pour un tableau contenant des variables qualitatives et 2 pour un tableau de fréquences.

Filtrer les facteurs : vous pouvez activer l'une ou les deux options suivantes afin de réduire le nombre de facteurs pour lesquels les résultats sont affichés :

- **% minimum** : activez cette option puis saisissez le pourcentage minimum de la variabilité totale que doivent représenter les facteurs retenus.

- **Nombre maximum** : activez cette option pour fixer le nombre maximum de facteurs à prendre en compte.

Options des ACP : (uniquement pour les tableaux de variables quantitatives)

- **Type d'ACP** : choisissez le type d'ACP à réaliser sur les tableaux : corrélation (les données seront centrées et réduites) ou covariance (les données seront uniquement centrées).

Options des ACM : (uniquement pour les tableaux de variables qualitatives)

Tri alphabétique des modalités : activez cette option pour que dans les divers résultats, les modalités soient triées alphabétiquement pour chacune des variables.

Libellés Variable-Modalité : activez cette option pour utiliser des libellés longs pour l'affichage des résultats. Les libellés Variable-Modalité sont composés du nom de la variable comme préfixe, et de la modalité comme suffixe.

Onglet **Données supplémentaires** :

Observations supplémentaires : activez cette option si vous voulez représenter des individus supplémentaires par le calcul de leurs coordonnées. Ces individus ne sont pas pris en compte pour le calcul des axes factoriels (observations passives, par opposition à observations actives). Si des libellés de variables sont présents pour les observations supplémentaires, vous devez activer l'option « Libellés des variables pour les obs. supp. ». Vous pouvez également choisir des libellés des observations supplémentaires pour l'affichage des résultats.

Tableaux supplémentaires : activez cette option si vous voulez utiliser certains tableaux comme tableaux illustratifs. Les variables de ces tableaux ne sont alors pas prises en compte pour le calcul des axes factoriels de l'AFM. Les analyses séparées sont en revanche effectuées pour les tableaux supplémentaires. Sélectionnez la plage des tableaux supplémentaires afin d'indiquer à XLSTAT quels sont, parmi les K tableaux, ceux qui sont actifs (1) ou illustratifs (0).

Onglet **Données manquantes** :

Ne pas accepter les valeurs manquantes : activez cette option pour que XLSTAT empêche la poursuite des calculs si des valeurs manquantes sont détectées.

Supprimer les observations : activez cette option pour supprimer les observations comportant des données manquantes.

Stratégies adaptées : activez cette option pour choisir des stratégies adaptées au type des données.

- Variables quantitatives :

- **Remplacer par la moyenne** : activez cette option pour estimer les données manquantes en utilisant la moyenne.

- **Plus proche voisin** : activez cette option pour estimer les données manquantes d'une observation en recherchant le plus proche voisin de l'observation.
- Variables qualitatives :
- **Nouvelle modalité** : une nouvelle catégorie « Manquant » est créée pour les variables qualitatives comprenant des valeurs manquantes.
- **Plus proche voisin** : activez cette option pour estimer les données manquantes d'une observation en recherchant le plus proche voisin de l'observation.

Onglet **Sorties** :

L'onglet Sorties est subdivisé en plusieurs sous-onglets :

Général :

Ces sorties concernent toutes les analyses :

Statistiques descriptives : activez cette option pour afficher les statistiques descriptives pour l'ensemble des variables sélectionnées.

Afficher les analyses séparées activez cette option pour afficher les résultats des analyses de chaque tableau pris séparément. Si l'option est désactivée, seuls les résultats de l'AFM seront affichés.

Valeurs propres : activez cette option pour afficher les tableaux et les graphiques (*scree plot*) des valeurs propres.

Contributions : activez cette option pour afficher les tableaux des contributions.

Cosinus carrés : activez cette option pour afficher les tableaux des cosinus carrés.

ACP :

Ces sorties concernent uniquement les ACP :

Coordonnées des variables : activez cette option pour afficher les coordonnées des variables dans l'espace des facteurs (*factor loadings* en anglais).

Corrélations Variables/Facteurs : activez cette option pour afficher les corrélations entre les facteurs et les variables.

Coordonnées des observations : activez cette option pour afficher les coordonnées des observations (*factor scores* en anglais) dans le nouvel espace créé par l'ACP.

ACM :

Ces sorties concernent uniquement les ACM :

Coordonnées des variables : activez cette option pour afficher les coordonnées des modalités des variables qualitatives dans l'espace des facteurs.

Coordonnées des observations : activez cette option pour afficher les coordonnées des observations dans l'espace des facteurs.

AFC :

Ces sorties concernent uniquement les AFC :

Coordonnées des colonnes : activez cette option pour afficher les coordonnées des colonnes dans l'espace des facteurs.

Coordonnées des lignes : activez cette option pour afficher les coordonnées des lignes dans l'espace des facteurs.

AFM :

Ces sorties concernent uniquement les résultats de seconde phase de l'AFM :

Tableaux :

- **Coordonnées** : activez cette option pour afficher les coordonnées des tableaux dans l'espace résultant de l'AFM. Remarque : les contributions et les cosinus sont aussi affichés si les options correspondantes ont été activées dans l'onglet Sorties/Général.
- **Coefficients Lg** : activez cette option pour afficher les coefficients Lg de liaison entre les tableaux.
- **Coefficients RV** : activez cette option pour afficher les coefficients RV de liaison entre les tableaux.

Variables :

- **Coordonnées des variables** : activez cette option pour afficher les coordonnées des variables dans l'espace résultant de l'AFM.
- **Corrélations Variables/Facteurs** : activez cette option pour afficher les corrélations entre les facteurs principaux et les variables.

Axes partiels :

- **Nombre maximum** : entrez le nombre maximum de facteurs à retenir des analyses de la première phase, que vous voulez ensuite analyser dans l'espace de l'AFM.
- **Coordonnées** : activez cette option pour afficher les coordonnées des axes partiels dans l'espace résultant de l'AFM.

- **Corrélations** : activez cette option pour afficher les corrélations entre les facteurs principaux et les axes partiels.
- **Corrélations entre les axes** : activez cette option pour afficher les corrélations entre les axes partiels.

Observations :

- **Coordonnées des observations** : activez cette option pour afficher les coordonnées des observations dans le nouvel espace créé par l'AFM.
- **Coordonnées des nuages partiels** : activez cette option pour afficher les coordonnées des nuages partiels dans l'espace résultant de l'AFM. Les nuages partiels correspondent aux projections des observations dans des espaces réduits aux dimensions de chacun des tableaux.

Onglet **Graphiques** :

L'onglet Graphiques est subdivisé en plusieurs sous-onglets :

Général :

Ces options concernent toutes les analyses :

Graphiques sur deux axes : activez cette option si vous souhaitez que les différentes représentations graphiques des ACP, ACM, AFC et AFM ne soient affichées que sur les deux premiers axes.

Options pour les variables :

Filtrer : activez cette option pour filtrer les variables affichées :

- **Aléatoire** : les variables à afficher sont sélectionnées de manière aléatoire. Le « Nombre de variables » doit alors être saisi.
- N **premières variables** : les N premières variables sont affichées. Le « Nombre de variables » N doit alors être saisi.
- N **dernières variables** : les N dernières variables sont affichées. Le « Nombre de variables » N doit alors être saisi.
- **Variable de groupe** : si vous choisissez cette option, vous devez ensuite sélectionner une variable indicatrice composée de 1 pour les variables à afficher, et de 0 pour les variables à ne pas afficher. Dans le cas de l'affichage des analyses séparées, cette option ne fonctionne pas.
- **Somme(Cos2)>** : choisissez cette option pour que, seules les variables ayant une somme des cosinus carrés supérieure à une valeur à saisir entre 0 et 1, soient affichées sur les graphiques de représentation des variables.

Options pour les observations :

Filtrer : activez cette option pour fixer le nombre d'observations affichées :

- **Aléatoire** : les observations à afficher sont sélectionnées de manière aléatoire. Le « Nombre d'observations » doit alors être saisi.
- N **premières lignes** : les N premières observations sont affichées. Le « Nombre d'observations » N doit alors être saisi.
- N **dernières lignes** : les N dernières observations sont affichées. Le « Nombre d'observations » N doit alors être saisi.
- **Variable de groupe** : si vous choisissez cette option, vous devez ensuite sélectionner une variable indicatrice composée de 1 pour les observations à afficher, et de 0 pour les observations à ne pas afficher.

Colorer par groupe : activez cette option si vous souhaitez colorer les observations en fonction d'une variable de groupe, sélectionnez un vecteur colonne de taille égale au nombre d'observations actives. Si des en-têtes de colonnes ont été sélectionnés pour le tableau principal, veillez à ce qu'un libellé soit aussi présent pour la variable de cette sélection.

```
* **Ellipses de confiance** : activez cette option si vous souhaitez afficher de
```

ACP :

Ces options concernent uniquement les graphiques des ACP :

Graphiques de corrélations : activez cette option pour afficher les graphiques mettant en jeu des corrélations entre des composantes et des variables initiales.

- **Vecteurs** : activez cette option pour afficher les variables d'origine sous forme de vecteurs.

Graphiques des observations : activez cette option pour afficher les graphiques de représentation des observations dans le nouvel espace.

- **Etiquettes** : activez cette option pour afficher les étiquettes des observations sur les graphiques. Le nombre d'étiquettes affichées peut être modulé à l'aide de l'option de filtrage.

Biplots : activez cette option pour afficher les graphiques de représentation simultanée des observations et des variables d'origine dans le nouvel espace.

- **Vecteurs** : activez cette option pour afficher les variables d'origine sous forme de vecteurs.
- **Etiquettes** : activez cette option pour afficher les étiquettes des observations sur les biplots. Le nombre d'étiquettes affichées peut être modulé à l'aide de l'option de filtrage.

Type de biplots : choisissez le type de biplot que vous souhaitez afficher. Voir la section description de l'ACP pour plus de détails.

- **Biplot de corrélation** : activez cette option pour afficher des biplots de corrélation.
- **Biplot de distance** : activez cette option pour afficher des biplots de distance.
- **Biplot symétrique** : activez cette option pour afficher des biplots symétriques.
- **Coefficient** : choisissez le coefficient dont la racine carrée sera multipliée par les coordonnées des variables. Ce coefficient vous permettra d'ajuster la position des points variables dans le biplot afin de rendre ce dernier plus lisible. Si ce coefficient est différent de 1, la longueur des vecteurs variables n'est plus interprétable en termes d'écart-type (biplot de corrélation) ou de contribution (biplot de distance).

ACM :

Ces options concernent uniquement les graphiques des ACM :

Carte factorielle des modalités : activez cette option pour afficher le graphique des coordonnées principales des modalités des variables qualitatives actives et supplémentaires.

- **Etiquettes** : activez cette option pour que les étiquettes des noms des modalités soient affichées sur le graphique.

Carte factorielle des observations : activez cette option pour afficher le graphique des coordonnées principales des observations actives et des observations supplémentaires.

- **Etiquettes** : activez cette option pour afficher les étiquettes des observations à côté des points.

Biplots : activez cette option pour afficher les graphiques de représentation simultanée des observations et des modalités.

- **Graphique symétrique** : choisissez cette option pour représenter sur un même graphique les coordonnées principales des modalités et les coordonnées principales des observations sur le même graphique.

AFC :

Ces options concernent uniquement les graphiques des AFC :

Graphique des colonnes : activez cette option pour afficher le graphique des coordonnées principales des colonnes sur le plan factoriel.

Graphique des lignes : activez cette option pour afficher le graphique des coordonnées principales des lignes sur le plan factoriel.

- **Etiquettes** : activez cette option pour afficher les étiquettes des observations à côté des points.

Biplots : activez cette option pour afficher les coordonnées principales des lignes et des colonnes sur un même graphique.

AFM :

Ces options concernent uniquement les graphiques de seconde phase de l'AFM :

Graphiques des tableaux : activez cette option pour afficher les graphiques de représentation des tableaux pour les différents axes choisis.

Graphiques des corrélations : activez cette option pour afficher le cercle des corrélations pour les variables quantitatives utilisées pour l'AFM.

Graphiques des observations : activez cette option pour afficher le cercle des corrélations pour les variables quantitatives utilisées pour l'AFM.

Graphiques de corrélations (axes partiels) : activez cette option pour afficher le graphique des observations dans l'espace de l'AFM.

Graphiques des nuages partiels : activez cette option pour afficher le graphique représentant à la fois les observations, et les observations projetés dans le sous-espace de chacun des tableaux.

- Libellés des observations : activez cette option pour afficher les libellés des observations sur les graphiques.
- Libellés des nuages partiels : activez cette option pour afficher les libellés des points des nuages partiels.

Résultats

Statistiques descriptives : le tableau de statistiques descriptives présente pour toutes les variables sélectionnées des statistiques simples. Sont affichés, le nombre d'observations, le nombre de données manquantes, le nombre de données non manquantes, la moyenne, et l'écart-type (non biaisé).

Sont ensuite affichés pour chacun des tableaux, les analyses séparées, si vous avez choisi d'afficher les analyses séparées.

A la suite des résultats des analyses séparées sont affichés les résultats de seconde phase de l'AFM.

Dans un premier temps les valeurs propres, les résultats des variables et les résultats des observations sont affichés comme dans le cas de l'ACP.

A la fin du tableau des coordonnées des observations, vous trouverez le bouton suivant :  .

Ce bouton vous permet d'ouvrir automatiquement la boite de dialogue pré-remplie de la CAH (Classification Ascendante Hiérarchique) afin d'effectuer une classification sur les coordonnées factorielles des observations.

Ensuite les résultats spécifiques à l'AFM sont affichés :

Les **coordonnées des tableaux** sont affichées et utilisées pour créer les graphiques des tableaux. Ces derniers permettent notamment de visualiser la distance entre les tableaux. Les

coordonnées des tableaux supplémentaires sont affichées dans la seconde partie du tableau. Ensuite, comme pour les variables et les observations sont affichés les contributions des tableaux actifs et les cosinus carrés des tableaux.

Coefficients Lg : les coefficients Lg de liaison entre les tableaux permettent de mesurer à quel point les tableaux sont liés deux à deux. La liaison sera d'autant plus forte que l'ensemble des variables d'un tableau seront liées à celle du second.

Coefficients RV : les coefficients RV de liaison entre les tableaux sont une autre mesure de la liaison entre les tableaux. Les coefficients RV dont la valeur est comprise entre 0 et 1, correspondent à une normalisation des coefficients Lg.

Les **coordonnées des axes partiels** et notamment leurs corrélations permettent de visualiser dans le nouvel espace le lien entre les facteurs générés par les analyses de la première phase de l'AFM (analyse des tableaux pris séparément), et ceux de la seconde étape (analyse de tous les tableaux pondérés).

Les **corrélations entre les axes partiels** permettent de voir les liaisons entre les axes des différentes analyses séparées.

Enfin, les **coordonnées des nuages partiels** dans l'espace résultant de l'AFM sont affichées. Les nuages partiels correspondent aux projections des observations dans des espaces réduits aux dimensions de chacun des tableaux. La représentation des points des nuages partiels superposée avec celles des observations complètes permet de visualiser à la fois la diversité de l'information apportée par les différents tableaux pour une observation donnée, et de visualiser les distances relatives de deux observations en fonction des différents tableaux.

*Remarque sur l'**indice d'homogénéité des axes*** : Cet indice développé par nos équipes est très utile pour déterminer si les contributions des observations sont homogènes pour les différents axes. Il est construit comme la proportion d'observations ayant une contribution absolue > 1/n. Un indice au-dessus de 0.4 indique une très bonne homogénéité avec des observations bien représentées. En revanche, un indice inférieur à 0.1 doit être une alerte pour l'utilisateur qui devrait vérifier s'il n'a pas de valeurs extrêmes sur les variables construisant l'axe qui fausseraient son interprétation (les valeurs extrêmes seraient alors les observations se démarquant des autres sur l'axe en question).

Exemple

Un exemple d'Analyse Factorielle Multiple sur des tableaux de différents types est disponible sur le Centre d'aide XLSTAT à l'adresse suivante :

http://www.xlstat.com/demo-mfaf.htm

Un exemple d'Analyse Factorielle Multiple sur des tableaux de fréquences est disponible sur le Centre d'aide XLSTAT à l'adresse suivante :

http://www.xlstat.com/demo-mfafreqf.htm

Bibliographie

Bécue-Bertaut M, Pagès J. (2008). Multiple factor analysis and clustering of a mixture of quantitative, categorical and frequency data, *Computational Statistics and Data Analysis* **vol. 52**

(pg. 3255-68).

Escofier B. and Pagès J. (1984). L'analyse factorielle multiple : une méthode de comparaison de groupes de variables. In : Sokal R.R., Diday E., Escoufier Y., Lebart L., Pagès J. (Eds), Data Analysis and Informatics III, 41-55. North-Holland, Amsterdam.

Escofier B. and Pagès J. (1994). Multiple Factor Analysis (AFMULT package). *Computational Statistics and Data Analysis*, **18**, 121-140.

Escofier B. and Pagès J. (1998). Analyses Factorielles Simples et Multiples : Objectifs, Méthodes et Interprétation. Dunod, Paris.

Robert P. and Escoufier Y. (1976). An unifying tool for linear multivariate methods. The RV coefficient. *Applied Statistics*, **25** (3), 257-265.

STATIS

Utilisez STATIS pour analyser plusieurs configurations objets/variables quantitatives. Cette méthode permet :

- d'étudier et visualiser les liens entres les objets;
- d'étudier les accords entre les configurations.

Dans cette section :

Description

La méthode STATIS est l'une des méthodes d'analyse de données multi configurations les plus utilisées en sensométrie. Les configurations sont alors les différents assesseurs, sujets, juges... Cette méthode peut notamment être utilisée dans le cas de données de projective mapping/Napping, profil conventionnel, profil libre... Le grand intérêt de STATIS réside dans le fait que les configurations atypiques ont un poids plus faible que celles qui sont centrales à l'ensemble des configurations. L'analyse reflète donc au mieux le point de vue général et non celui des configurations atypiques.

Utilisations de STATIS

Il existe plusieurs applications pour STATIS, parmi lesquelles :

- Etude et visualisation des objets dans les plans principaux;
- Etude des liens entre les configurations, notamment pour trouver les plus atypiques.

Principe de STATIS

STATIS est une méthode travaillant sur la matrice des produits scalaires de chaque configuration, ce qui permet de travailler avec des configurations ayant des nombres de colonnes différents. Son objectif est de former une configuration consensus qui reflète au mieux les différentes configurations. Ce consensus peut ensuite être projeté sur différents axes. Si l'information associée aux 2 ou 3 premiers axes représente un pourcentage suffisant de la

variabilité totale du consensus, on pourra représenter les objets sur un graphique à 2 ou 3 dimensions, facilitant ainsi grandement l'interprétation.

Structure des données

Il existe deux cas différents :

1. Le nombre de variables est identique pour les m configurations.

2. Le nombre de variables varie d'une configuration à l'autre.

Pour la saisie des données, XLSTAT vous demande de sélectionner une configuration correspondant aux m configurations contiguës, et de donner le cas de structure. Si les données ne sont **pas adaptées** à ce type de tableau, il est possible de les **transformer** en un tableau horizontal en utilisant la fonctionnalité **Créer un tableau Produits\Sujets.**

Mise à l'échelle et réduction globale

Si les données au sein même d'une configuration ne sont pas à la même échelle, et uniquement dans ce cas, il est conseillé de mettre à l'échelle (réduire) les variables de chaque configuration. Ce n'est par exemple pas le cas pour des notes attribuées entre 0 et 20 pour différents attributs, mais conseillé si certaines notes sont entre 0 et 10 et d'autres entre 0 et 20.

Classiquement, la réduction globale de chaque configuration est conseillée. Elle permet de mettre toutes les configurations sur un pied d'égalité en terme de variance. Dans le cas par exemple de configurations où les attributs sont notés entre 0 et 20 par des assesseurs, elle permettra d'enlever les facteurs d'échelle entre l'assesseur qui note uniquement entre 5 et 15 et l'assesseur qui utilise toute la gamme de notes.

Interprétation des résultats

La représentation des objets dans l'espace des k facteurs permet d'interpréter visuellement les proximités entre les objets, moyennant certaines précautions.

On peut considérer que la projection d'un objet sur un plan est fiable si l'objet est éloigné du centre du graphique.

Nombre de facteurs

Deux méthodes sont communément utilisées pour déterminer quel nombre de facteurs doit être retenu pour l'interprétation des résultats :

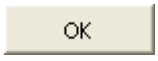
- Regarder la courbe décroissante des valeurs propres. Le nombre de facteurs à retenir correspond au premier point d'inflexion sur la courbe.
- On peut aussi se fonder sur le pourcentage cumulé de variabilité représenté par les axes factoriels et décider de se contenter d'un certain pourcentage.

Représentations graphiques

Les représentations graphiques ne sont fiables que si la somme des pourcentages de variabilité associés aux axes de l'espace de représentation, est suffisamment élevée. Si ce pourcentage est élevé (par exemple 80%), on peut considérer que la représentation est fiable. Si le pourcentage est faible, il est conseillé de faire des représentations sur plusieurs paires d'axes afin de valider l'interprétation faite sur les deux premiers axes factoriels.

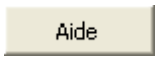
Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.



: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.

Onglet **Général** :

Configurations : sélectionnez les données correspondant aux configurations. Si la première ligne de la sélection comprend des en-têtes, l'option « Libellés des variables » doit être activée. Si les données ne sont **pas adaptées** à ce type de tableau, il est possible de les **transformer** en un tableau horizontal en utilisant la fonctionnalité **Créer un tableau Produits\Sujets**.

Nombre de configurations : entrez le nombre de configurations contenues dans la sélection ci-dessus.

Nombre de variables par configuration :

- **Egal** : choisissez cette option si le nombre de variables est identique pour toutes les configurations. XLSTAT détermine alors automatiquement le nombre de variables de chacune des configurations.

- **Défini par l'utilisateur** : choisissez cette option pour sélectionner une plage contenant les nombres de variables correspondant à chacune des configurations. Si l'option « Libellés des variables » est activée, la première cellule de la sélection doit comprendre un en-tête.

Libellés des configurations : activez cette option si vous voulez utiliser les libellés des configurations pour l'affichage des résultats. Si l'option « Libellés des variables » est activée, la première cellule de la sélection doit comprendre un en-tête. Si vous n'activez pas cette option, des libellés seront automatiquement créés (Config.1, Config.2, …).

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des variables : activez cette option si la première ligne des données sélectionnées (configurations, libellés des configurations, libellés des objets) contient un libellé.

Libellés des objets : activez cette option si vous voulez utiliser des libellés d'objets pour l'affichage des résultats. Si l'option « Libellés des variables » est activée, la première cellule de la sélection doit comprendre un en-tête. Si vous n'activez pas cette option, des libellés seront automatiquement créés (Obj.1, Obj.2, …).

Onglet **Options** :

Mise à l'échelle (variables) : activez cette option pour réduire les variables, c'est à dire les mettre à la même échelle.

Réduction globale : activez cette option pour réduire globalement les configurations.

Filtrer les facteurs : vous pouvez activer l'une des deux options suivantes afin de réduire le nombre de facteurs pour lesquels les résultats sont affichés :

- **% minimum** : activez cette option puis saisissez le pourcentage minimum de la variabilité totale que doivent représenter les facteurs retenus.
- **Nombre maximum** : activez cette option pour fixer le nombre maximum de facteurs à prendre en compte.

Onglet **Données manquantes** :

Ne pas accepter les valeurs manquantes : activez cette option pour que XLSTAT empêche la poursuite des calculs si des valeurs manquantes sont détectées.

Supprimer les observations : activez cette option pour supprimer les observations comportant des données manquantes.

Estimer les données manquantes : activez cette option pour estimer les données manquantes avant le début des calculs.

- **Moyenne ou mode** : activez cette option pour estimer les données manquantes en utilisant la moyenne (variables quantitatives) ou le mode (variables qualitatives) pour les variables correspondantes.
- **Plus proche voisin** : activez cette option pour estimer les données manquantes d'une observation en recherchant le plus proche voisin de l'observation.

Onglet **Sorties** :

Statistiques descriptives : activez cette option pour afficher les statistiques descriptives pour les variables sélectionnées.

Valeurs propres : activez cette option pour afficher le tableau des valeurs propres.

Coordonnées du consensus : activez cette option pour afficher les coordonnées du consensus dans l'espace des facteurs.

Matrice RV : activez cette option pour afficher la matrice des coefficients RV.

Facteurs de mise à l'échelle : activez cette option pour afficher les facteurs de mise à l'échelle des configurations.

Poids : activez cette option pour afficher les poids créés et utilisés par STATIS.

Configuration consensus : activez cette option pour afficher la configuration consensus créée par STATIS.

Homogénéité : activez cette option pour afficher l'homogénéité des configurations.

RV config/consensus : activez cette option pour afficher le coefficient RV entre chaque configuration et le consensus.

Erreur globale : activez cette option pour afficher l'erreur du critère STATIS.

Résidus par configuration : activez cette option pour afficher les résidus de chaque configuration du critère STATIS.

Résidus par objet : activez cette option pour afficher les résidus de chaque objet du critère STATIS.

Corrélations : activez cette option pour afficher les corrélations entre les facteurs et les variables initiales.

Coordonnées des nuages partiels : activez cette option pour afficher les coordonnées des nuages partiels dans l'espace des facteurs. Les nuages partiels correspondent aux projections des objets de chacune des configurations dans l'espace des facteurs.

- **Présentation par configuration** : activez cette option pour afficher un tableau de coordonnées par configuration.
- **Présentation par objet** : activez cette option pour afficher un tableau de coordonnées par objet.

Onglet **Graphiques** :

Graphiques sur deux axes : activez cette option si vous souhaitez que les différentes représentations graphiques ne soient affichées que sur les deux premiers axes.

Valeurs propres : activez cette option pour afficher le graphique (*scree plot*) des valeurs propres.

Coordonnées du consensus : activez cette option pour afficher le graphique des coordonnées du consensus dans l'espace des facteurs.

Facteurs de mise à l'échelle : activez cette option pour afficher le diagramme en bâtons des facteurs de mise à l'échelle des configurations.

Poids : activez cette option pour afficher le diagramme en bâtons des poids créés et utilisés par STATIS.

RV config/consensus : activez cette option pour afficher le diagramme en bâtons des coefficients RV entre chaque configuration et le consensus.

Résidus par configuration : activez cette option pour afficher le diagramme en bâtons des résidus de chaque configuration du critère STATIS.

Résidus par objet : activez cette option pour afficher le diagramme en bâtons des résidus de chaque objet du critère STATIS.

Corrélations : activez cette option pour afficher les graphiques mettant en jeu des corrélations entre les facteurs et les variables initiales. Ce graphique est communément appelé cercle des corrélations.

Graphiques des nuages partiels : activez cette option pour afficher les graphiques représentant à la fois les objets, et les objets de chacun des tableaux projetés dans l'espace des facteurs.

- Libellés des observations : activez cette option pour afficher les libellés des observations sur les graphiques.
- Libellés des nuages partiels : activez cette option pour afficher les libellés des points des nuages partiels.

Résultats

Statistiques descriptives : le tableau de statistiques descriptives présente pour toutes les variables sélectionnées des statistiques simples. Sont affichés le nombre d'observations, le

nombre de données manquantes, le nombre de données non manquantes, la moyenne, et l'écart-type (non biaisé).

Valeurs propres : les valeurs propres et le graphique (*scree plot*) correspondant sont affichés.

Coordonnées du consensus : les coordonnées du consensus dans l'espace des facteurs sont affichées, ainsi que les graphiques correspondants (en fonction du nombre de facteurs choisi).

Matrice RV : la matrice des coefficients RV entre toutes les configurations est affichée. Le coefficient RV est un coefficient de similarité entre deux configurations compris entre 0 et 1. Plus il est proche de 1, plus la similarité est forte. Cette matrice est utilisée par STATIS pour calculer les poids des configurations.

Facteurs de mise à l'échelle : les facteurs d'échelle sont affichés, ainsi que le diagramme en bâtons associé. Plus un facteur d'échelle d'une configuration est grand, plus l'échelle de la configuration utilisée est restreinte. Ce tableau est utilisé en analyse sensorielle pour comprendre comment les assesseurs utilisent différemment les échelles de notation.

Poids : les poids calculés par STATIS sont affichés, ainsi que le diagramme en bâtons associé. Plus un poids est grand, plus la configuration a contribué à l'élaboration du consensus. Sachant que STATIS donne du poids aux configurations les plus proches du point de vue global, un poids beaucoup plus faible que les autres signifiera que la configuration est atypique.

Configuration consensus : la configuration consensus créée par STATIS est affichée. Elle correspond à la moyenne pondérée par les poids des matrices de produits scalaires des configurations initiales (éventuellement réduites par variable et/ou globalement).

Homogénéité : l'homogénéité des configurations est affichée. C'est une valeur comprise entre $1/m$ (m étant le nombre de configurations) et 1, qui croît avec l'homogénéité des configurations.

RV config/consensus : les coefficients RV entre les configurations et le consensus sont affichés, ainsi que le diagramme en bâtons associé. Tout comme les poids de STATIS, ces coefficients permettent de détecter des configurations atypiques. L'avantage de ces coefficients est qu'ils sont compris entre 0 et 1, donc plus faciles à interpréter que les poids.

Erreur globale : l'erreur du critère STATIS est affichée. Elle correspond à la somme de tous les résidus (qui peuvent être présentés par configuration ou par objet).

Résidus par configuration : ce tableau et le diagramme en bâtons correspondant permettent de visualiser la répartition des résidus de STATIS par configuration. On peut ainsi repérer pour quelles configurations STATIS a été moins efficace, autrement dit, quelles configurations se démarquent le plus de la configuration consensus.

Résidus par objet : ce tableau et le diagramme en bâtons correspondant permettent de visualiser la répartition des résidus de STATIS par objet. On peut ainsi repérer pour quels objets STATIS a été moins efficace, autrement dit, quels objets se démarquent le plus de la configuration consensus.

Corrélations : les corrélations entre les facteurs et les variables initiales, ainsi que le cercle des corrélations associé sont affichés. Ce graphique permet de voir les liens entre les différentes variables et les facteurs.

Coordonnées des objets (présentation par configuration) : cette série de tableaux correspond aux coordonnées des objets pour chaque configuration, après les éventuelles mises à l'échelle et réductions globales puis la projection sur les facteurs. La présentation est faite par configuration.

Coordonnées des objets (présentation par objet) : cette série de tableaux correspond aux coordonnées des objets pour chaque configuration, après les éventuelles mises à l'échelle et réductions globales puis la projection sur les facteurs. La présentation est faite par objet.

Coordonnées des nuages partiels : Les nuages partiels correspondent aux projections des objets de chacune des configurations dans l'espace des facteurs. La représentation des points des nuages partiels superposée avec celles des objets permet de visualiser à la fois la diversité de l'information apportée par les différentes configurations pour un objet donné, et de visualiser les distances relatives de deux objets en fonction des différentes configurations.

Exemple

Un exemple d'utilisation de STATIS est disponible sur le Centre d'aide XLSTAT à l'adresse suivante :

http://www.xlstat.com/demo-stifr.htm

Bibliographie

Lavit, C., Escoufier, Y., Sabatier, R., Traissac, P. (1994). The ACT (STATIS method). *Computational Statistics & Data Analysis*, **18** , 1, 97-119.

Llobell, F., Cariou, V., Vigneau, E., Labenne, A., Qannari, E. M. (2018). Analysis and clustering of multiblock datasets by means of the STATIS and CLUSTATIS methods. Application to sensometrics. *Food Quality and Preference*, **79**, 103520.

Llobell, F. (2020). Classification de tableaux de données, applications en analyse sensorielle (Doctoral dissertation, Nantes, Ecole nationale vétérinaire).

Schlich, P. (1996). Defining and validating assessor compromises about product distances and attribute Correlations. *In: Multivariate Analysis of Data in Sensory Science*, 259-306.

CLUSTATIS

Utilisez CLUSTATIS pour constituer des classes homogènes de configurations/tableaux de données. Dans le cadre de l'analyse sensorielle, cette fonction permet de réaliser une classification des sujets sur la base de leurs perceptions des produits.

Dans cette section :

Description

Les cas où les données sont constituées de différents blocs de variables sont de plus en plus fréquents. L'analyse sensorielle est particulièrement concernée par ce phénomène, puisque de nombreuses épreuves amènent à ce type de données, chaque consommateur/juge/sujet apportant un tableau (une configuration) de données (ex: épreuve de Projective mapping/Napping, profil conventionnel, profil libre). Étant donné que les perceptions entre les sujets sont bien souvent différentes, une classification de ces derniers peut s'avérer nécessaire. La méthode CLUSTATIS s'inscrit dans ce contexte. De plus, cette stratégie permet de mettre de côté les configurations qui ne se conforment à aucune des classes construites, qui correspondent à des sujets atypiques dans le cadre de l'analyse sensorielle.

Principe de CLUSTATIS

CLUSTATIS est une méthode de classification basée sur les matrices des produits scalaires de chaque configuration, ce qui permet de considérer des configurations ayant des nombres de colonnes différents. L'objectif de cette méthode est de constituer des classes de configurations les plus homogènes possible, chaque groupe de configurations étant représenté par une configuration latente (nommée consensus) déterminée par STATIS. Il est donc naturel que chaque classe soit finalement analysée par STATIS, afin de déterminer les différences entre les classes constituées. CLUSTATIS consiste en un algorithme hiérarchique pouvant être « consolidé » par un algorithme de partitionnement (c'est à dire que l'algorithme de partitionnement est initialisé par la coupe du dendrogramme). Une option intéressante est la création d'une classe « K+1 » (correspondant à une classe supplémentaire) afin de mettre de côté les tableaux ne se conformant à aucune classe. Une configuration sera placée dans cette classe si les similarités (coefficents RV) entre le consensus de chaque classe et cette configuration sont tous considérés comme faibles.

Structure des données

Il existe deux cas différents :

1. Le nombre de variables est identique pour les m configurations.

2. Le nombre de variables varie d'une configuration à l'autre.

Pour la saisie des données, XLSTAT vous demande de sélectionner une configuration correspondant aux m configurations contiguës, et de donner le cas de structure. Si les données ne sont **pas adaptées** à ce type de tableau, il est possible de les **transformer** en un tableau horizontal en utilisant la fonctionnalité **Créer un tableau Produits\Sujets.**

Mise à l'échelle

Si les données au sein même d'une même configuration ne sont pas à la même échelle, et uniquement dans ce cas, il est conseillé de mettre à l'échelle (réduire) les variables de chaque configuration. Ce n'est par exemple pas le cas pour des notes attribuées entre 0 et 20 pour différents attributs sensoriels, mais conseillé si certaines notes sont entre 0 et 10 et d'autres entre 0 et 20.

Interprétation des résultats

Pour chaque classe, la représentation des objets/observations dans l'espace des facteurs permet d'interpréter visuellement les proximités entre ces objets, moyennant certaines précautions. On peut considérer que la projection d'un objet sur un plan est fiable si l'objet est éloigné du centre du graphique.

Étant donné que la classe « K+1 » contient les tableaux ne se conformant à aucune des classes, cette classe est très dépendante du nombre de groupes.

Nombre de facteurs

Deux méthodes sont communément utilisées pour déterminer le nombre de facteurs à retenir pour l'interprétation des résultats :

- Regarder la courbe décroissante des valeurs propres. Le nombre de facteurs à retenir correspond au premier point d'inflexion sur la courbe.
- On peut aussi se baser sur le pourcentage cumulé de variabilité représenté par les axes factoriels et décider de se contenter d'un certain pourcentage.

Représentations graphiques

Les représentations graphiques ne sont fiables que si la somme des pourcentages de variabilité associé aux axes de l'espace de représentation est suffisamment élevée. Si ce pourcentage est élevé (par exemple 80%), on peut considérer que la représentation est fiable. Si le pourcentage est faible, il est conseillé de faire des représentations sur plusieurs paires d'axes afin de valider l'interprétation faite sur les deux premiers axes factoriels.

Qualité de la classification

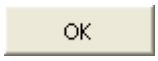
Afin de déterminer la qualité d'une classification hiérarchique, on peut s'aider de l'augmentation de la variance intra-classes (erreur du critère CLUSTATIS) provoquée par la fusion de deux classes. Cette augmentation est égale à la hauteur du dendrogramme à laquelle les deux classes de configurations se retrouvent rassemblées dans la même classe.

L'homogénéité de chaque classe et l'homogénéité globale sont également des indices très importants (entre $1/m$ et 1, m étant le nombre de configurations) qui permettent de juger de la

qualité de la classification. Il est à noter que la consolidation et l'ajout d'une classe « K+1 » peuvent augmenter les homogénéités.

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.


: cliquez sur ce bouton pour changer la façon dont vous chargez les données dans XLSTAT. Si le bouton possède une icône de souris, XLSTAT vous permet de faire une sélection à la souris des données. Si le bouton possède une icône de liste, XLSTAT vous permet de sélectionner les données à partir d'une liste. Si le bouton possède une icône orange, des boutons complémentaires avec un point d'interrogation  sont affichés pour vous permettre d'importer les données à partir de fichiers.

Onglet **Général** :

Configurations : sélectionnez les données correspondant aux configurations. Si la première ligne de la sélection comprend des en-têtes, l'option « Libellés des variables » doit être activée. Si les données ne sont **pas adaptées** à ce type de tableau, il est possible de les **transformer** en un tableau horizontal en utilisant la fonctionnalité **Créer un tableau Produits\Sujets**.

Nombre de configurations : entrez le nombre de configurations contenues dans la sélection ci-dessus.

Nombre de variables par configuration :

- **Egal** : choisissez cette option si le nombre de variables est identique pour toutes les configurations. XLSTAT détermine alors automatiquement le nombre de variables de chacune des configurations.

- **Défini par l'utilisateur** : choisissez cette option pour sélectionner une plage contenant les nombres de variables correspondant à chacune des configurations. Si l'option « Libellés des variables » est activée, la première cellule de la sélection doit comprendre un en-tête.

Libellés des configurations : activez cette option si vous voulez utiliser les libellés des configurations pour l'affichage des résultats. Si l'option « Libellés des variables » est activée, la première cellule de la sélection doit comprendre un en-tête. Si vous n'activez pas cette option, des libellés seront automatiquement créés (Config.1, Config.2, …).

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des variables : activez cette option si la première ligne des données sélectionnées (Configurations, Libellés des configurations, Libellés des objets) contient un libellé.

Libellés des objets : activez cette option si vous voulez utiliser des libellés d'objets pour l'affichage des résultats. Si l'option « Libellés des variables » est activée, la première cellule de la sélection doit comprendre un en-tête. Si vous n'activez pas cette option, des libellés seront automatiquement créés (Obj.1, Obj.2, …).

Onglet **Options** :

Mise à l'échelle (variables) : activez cette option pour réduire les variables, c'est à dire les mettre à la même échelle.

Filtrer les facteurs : vous pouvez activer l'une des deux options suivantes afin de réduire le nombre de facteurs pour lesquels les résultats sont affichés :

- **% minimum** : activez cette option puis saisissez le pourcentage minimum de la variabilité totale que doivent représenter les facteurs retenus.

- **Nombre maximum** : activez cette option pour fixer le nombre maximum de facteurs à prendre en compte.

Troncature : activez cette option si vous voulez que XLSTAT définisse **automatiquement** une troncature, et donc le nombre de classes à retenir, ou si vous voulez définir vous-même le **nombre de classes** à créer, ou si vous voulez définir le **niveau** auquel le dendrogramme doit être tronqué.

Consolidation : activez cette option pour réaliser une consolidation des classes obtenues à partir du dendrogramme.

Classe K+1 : activez cette option pour ajouter une classe supplémentaire qui contiendra les configurations ne se conformant à aucune des classes.

Paramètre rho: Choisissez la façon dont vous voulez définir le paramètre rho:**automatiquement** ou **défini par l'utilisateur**. Ce paramètre représente l'accord minimal pour être considéré comme suffisammment en accord pour être conservé dans une classe. Plus ce paramètre augmente, plus l'accord demandé avec la classe est fort et plus vous risquez de placer de configurations dans la classe K+1.

Onglet **Données manquantes** :

Ne pas accepter les valeurs manquantes : activez cette option pour que XLSTAT empêche la poursuite des calculs si des valeurs manquantes sont détectées.

Supprimer les observations : activez cette option pour supprimer les observations comportant des données manquantes.

Estimer les données manquantes : activez cette option pour estimer les données manquantes avant le début des calculs.

- **Moyenne ou mode** : activez cette option pour estimer les données manquantes en utilisant la moyenne (variables quantitatives) ou le mode (variables qualitatives) pour les variables correspondantes.
- **Plus proche voisin** : activez cette option pour estimer les données manquantes d'une observation en recherchant le plus proche voisin de l'observation.

Onglet **Sorties** :

Statistiques descriptives : activez cette option pour afficher les statistiques descriptives pour les variables sélectionnées.

Matrice RV : activez cette option pour afficher la matrice des coefficients RV entre les configurations.

Statistiques des nœuds : activez cette option pour afficher les statistiques des nœuds du dendrogramme.

Composition des classes : activez cette option pour afficher la composition de chacune des classes.

Valeurs propres : activez cette option pour afficher le tableau des valeurs propres de chaque classe.

Coordonnées du consensus : activez cette option pour afficher les coordonnées du consensus de chaque classe dans l'espace des facteurs.

Configuration consensus : activez cette option pour afficher la configuration consensus de chaque classe créée par STATIS.

RV config/consensus : activez cette option pour afficher le coefficient RV entre chaque configuration et le consensus de sa classe.

Poids : activez cette option pour afficher les poids créés et utilisés par STATIS dans chaque classe.

Homogénéités : activez cette option pour afficher l'homogénéité de chacune des classes ainsi que l'homogénéité globale.

Erreur globale/Variance intra-classes : activez cette option pour afficher l'erreur du critère de minimisation CLUSTATIS, équivalente à la variance intra-classes.

RV entre consensus : activez cette option pour afficher le coefficient RV entre chaque configuration consensus.

Onglet **Graphiques** :

Diagramme des niveaux : activez cette option pour afficher le diagramme des niveaux permettant d'observer l'impact des regroupements successifs sur la variance intra-classes.

Dendrogramme : activez cette option pour afficher le dendrogramme.

- **Horizontal** : choisissez cette option pour afficher un dendrogramme horizontal.
- **Vertical** : choisissez cette option pour afficher un dendrogramme vertical.
- **Tronqué** : activez cette option pour afficher le dendrogramme tronqué (le dendrogramme commence au niveau de la troncature).
- **Etiquettes** : activez cette option pour afficher les libellés des configurations (dendrogramme complet) ou des classes (dendrogramme tronqué) sur le dendrogramme.
- **Couleurs** : activez cette option pour utiliser des couleurs pour représenter les différents groupes sur le dendrogramme complet.

Graphiques sur deux axes : activez cette option si vous souhaitez que les différentes représentations graphiques ne soient affichées que sur les deux premiers axes.

Valeurs propres : activez cette option pour afficher le graphique (*scree plot*) des valeurs propres de chaque classe.

Coordonnées du consensus : activez cette option pour afficher le graphique des coordonnées du consensus de chaque classe dans l'espace des facteurs.

RV config/consensus : activez cette option pour afficher le diagramme en bâtons des coefficients RV entre chaque configuration et le consensus de sa classe.

Poids : activez cette option pour afficher le diagramme en bâtons des poids créés et utilisés par STATIS dans chaque classe.

Résultats

Statistiques descriptives : le tableau de statistiques descriptives présente pour toutes les variables sélectionnées des statistiques simples. Sont affichés le nombre d'observations, le

nombre de données manquantes, le nombre de données non manquantes, la moyenne, et l'écart-type (non biaisé).

Matrice RV : la matrice des coefficients RV entre toutes les configurations est affichée. Le coefficient RV est un indice de similarité entre deux configurations compris entre 0 et 1. Plus il est proche de 1, plus la similarité est forte.

Statistiques des nœuds : dans ce tableau sont affichées les informations concernant les nœuds successifs du dendrogramme. Le premier nœud a pour indice le nombre de configurations augmenté de 1. Ainsi, il est aisé de repérer à quel moment une configuration ou un groupe de configurations est regroupé avec un autre groupe de configurations dans le dendrogramme.

Diagramme des niveaux : dans ce graphique sont affichés les niveaux des nœuds du dendrogramme, qui correspondent à l'augmentation du critère de minimisation de CLUSTATIS (équivalent à l'augmentation de la variance intra-classes) lors de la fusion de deux classes.

Dendrogrammes : le dendrogramme complet permet de visualiser le regroupement progressif des configurations. Si une troncature a été demandée, un trait en pointillé marque le niveau auquel est effectuée la troncature. Le dendrogramme tronqué permet de visualiser les classes après la troncature.

Compositions des classes :

Résultats par configuration : dans ce tableau est indiquée pour chaque configuration sa classe d'affectation dans l'ordre initial des configurations. Si une consolidation est demandée, les résultats sont donnés avant et après la consolidation. Dans le cas où vous avez coché classe « K+1 », il est possible que certains tableaux aient une valeur manquante après la consolidation. Ceci signifie qu'ils ne sont placés dans aucune des classes principales (ils sont placés dans la classe « K+1 »).

Résultats par classe : Les résultats sont donnés par classe. Ainsi, une liste de configurations est affichée pour chacune des classes.

Nombre de configurations par classe : Le nombre de configurations dans chaque classe est indiqué.

Paramètre rho calculé : Résultat affiché uniquement si vous avez choisi d'ajouter une classe « K+1 ». Le paramètre rho représente la similarité minimale que doit avoir une configuration avec le consensus d'une classe pour lui appartenir. Si cette condition n'est repectée pour aucune des classes, la configuration est placée dans la classe « K+1 ». Ce paramètre est calculé en fonction de la proximité de chaque configuration avec sa classe ainsi qu'avec la classe voisine.

Analyse de la classe k :

Dans cette section est affichée l'analyse de chacune des classes par la méthode STATIS. Chaque classe est analysée tour à tour.

Valeurs propres : les valeurs propres et le graphique (*scree plot*) correspondant sont affichés.

Coordonnées du consensus : les coordonnées du consensus dans l'espace des facteurs sont affichées, ainsi que les graphiques correspondants (en fonction du nombre de facteurs choisi).

Configuration consensus : la configuration consensus créée par STATIS est affichée. Elle correspond à la moyenne pondérée par les poids des matrices de produits scalaires des configurations initiales (réduites globalement et éventuellement par variable).

RV config/consensus : les coefficients RV entre les configurations et le consensus sont affichés, ainsi que le diagramme en bâtons associé. Tout comme les poids de STATIS, ces coefficients permettent de détecter des configurations atypiques. L'avantage de ces coefficients est qu'ils sont compris entre 0 et 1, donc plus faciles à interpréter que les poids.

Poids : les poids calculés par STATIS sont affichés, ainsi que le diagramme en bâtons associé. Plus un poids est grand, plus la configuration a contribué à l'élaboration du consensus. Sachant que STATIS donne du poids aux configurations les plus proches du point de vue global, un poids beaucoup plus faible que les autres signifiera que la configuration est atypique.

Indices :

Homogénéités : l'homogénéité de chaque classe est affichée. C'est une valeur comprise entre $1/m$ (m étant le nombre de configurations de la classe) et 1, qui croît avec l'homogénéité des configurations. Dans un second temps, l'homogénéité globale, qui est une moyenne pondérée des homogénéités de chaque classe, est affichée.

Erreur globale/Variance intra-classes : l'erreur du critère CLUSTATIS est affichée. Elle correspond à la variance intra-classes.

RV entre consensus : la matrice des coefficients RV entre les consensus de chaque classe est affichée. Cette matrice permet de voir dans quelle mesure les classes sont proches les unes des autres.

Exemple

Un exemple d'utilisation de CLUSTATIS est disponible sur le Centre d'aide XLSTAT à l'adresse suivante :

http://www.xlstat.com/demo-cstf.htm

Bibliographie

Llobell, F., Cariou, V., Vigneau, E., Labenne, A., Qannari, E. M. (2020). Analysis and clustering of multiblock datasets by means of the STATIS and CLUSTATIS methods. Application to sensometrics. *Food Quality and Preference*, **79**, 103520.

Llobell, F., Vigneau, E., & Qannari, E. M. (2019). Clustering datasets by means of CLUSTATIS with identification of atypical datasets. Application to sensometrics. *Food quality and preference*, **75**, 97-104.

Llobell, F. (2020). Classification de tableaux de données, applications en analyse sensorielle (Doctoral dissertation, Nantes, Ecole nationale vétérinaire).

CATATIS

Utilisez CATATIS pour analyser vos données Check-All-That-Apply (CATA). Cette méthode permet :

- d'étudier et visualiser les liens entres les produits et les attributs;
- d'étudier les accords entre les sujets.

Dans cette section :

Description

La méthode CATATIS est une amélioration de la méthode usuelle pour traiter les données CATA. Elle est considérée comme l'équivalent de la méthode STATIS pour ce type de données. Le grand intérêt de CATATIS réside dans le fait que les sujets atypiques ont un poids plus faible que ceux qui sont en accord avec le reste du panel. L'analyse reflète donc au mieux le point de vue général et non celui des sujets atypiques.

De plus, des tests de consistance du panel globalement et par attribut sont proposés afin de déterminer si certains attributs ne sont pas compris. Des tests sur les poids pour déterminer si certains sujets ont un poids non significatif peuvent également être réalisés. Cette dernière option est particulièrement utile lorsque nous avons affaire à des experts.

Le fait que les sujets aient eu plusieurs sessions est autorisé. Dans ce cas, alors la moyenne est utilisée par sujet (si pour un produit et un attribut donné il a coché une fois et non coché l'autre fois, sa moyenne sera de 0.5). De plus, les données non binaires sont acceptées et la répétabilité d"une session à l'autre est controlée.

Utilisation de CATATIS

Il existe plusieurs applications pour CATATIS, parmi lesquelles :

- l'étude et la visualisation des produits et attributs dans les plans factoriels principaux ;
- l'étude de la similarité entre les sujets, notamment pour trouver les plus atypiques.

Principe de CATATIS

L'objectif de CATATIS est de former une configuration consensus qui reflète au mieux les différents sujets. Ce consensus peut ensuite être projeté sur différents axes factoriels à l'aide d'une Analyse Factorielle des correspondances (AFC). Si l'information associée aux 2 ou 3 premiers axes représente un pourcentage suffisant de la variabilité totale du consensus, on pourra représenter les produits et les attributs sur un graphique à 2 ou 3 dimensions, facilitant ainsi grandement l'interprétation.

Structure des données

Il existe deux formats différents :

1. Toutes les données sont concaténées horizontalement (format horizontal).
2. Toutes les données sont concaténées verticalement (format vertical).

Pour la saisie des données, XLSTAT vous demande de sélectionner l'ensemble des données, et de donner le type de format. Dans le cas du format vertical, les produits et les sujets sont demandés.

Interprétation des résultats

La représentation des produits et attributs dans l'espace des k facteurs permet d'interpréter visuellement les proximités entre les produits et les attributs, moyennant certaines précautions.

On peut considérer que la projection d'un produit ou d'un attribut sur un plan est fiable si elle est éloignée du centre du graphique.

Nombre de facteurs

Deux méthodes sont communément utilisées pour déterminer le nombre de facteurs à retenir pour l'interprétation des résultats :

- regarder la courbe décroissante des valeurs propres. Le nombre de facteurs à retenir correspond au premier point d'inflexion sur la courbe ;
- on peut aussi se baser sur le pourcentage cumulé de variabilité représenté par les axes factoriels et décider de se contenter d'un certain pourcentage.

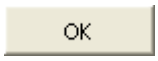
Représentations graphiques

Les représentations graphiques ne sont fiables que si la somme des pourcentages de variabilité associés aux axes de l'espace de représentation, est suffisamment élevée. Si ce pourcentage est élevé (par exemple 80%), on peut considérer que la représentation est fiable. Si le

pourcentage est faible, il est conseillé de faire des représentations sur plusieurs paires d'axes afin de valider l'interprétation faite sur les deux premiers axes factoriels.

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.


: cliquez sur ce bouton pour changer la façon dont vous chargez les données dans XLSTAT. Si le bouton possède une icône de souris, XLSTAT vous permet de faire une sélection à la souris des données. Si le bouton possède une icône de liste, XLSTAT vous permet de sélectionner les données à partir d'une liste. Si le bouton possède une icône orange, des boutons complémentaires avec un point d'interrogation  sont affichés pour vous permettre d'importer les données à partir de fichiers.

Onglet **Général** :

Données CATA (0/1) : sélectionnez les données correspondant aux différents sujets. Si la première ligne de la sélection comprend des en-têtes, l'option « Libellés des attributs » doit être activée.

Format : Cliquez sur horizontal ou vertical selon la façon dont sont structurées vos données.

Si le format est **horizontal** :

Nombre de sujets : entrez le nombre de sujets dans les données CATA (format horizontal uniquement).

Libellés des produits : activez cette option si vous voulez utiliser des libellés des produits pour l'affichage des résultats. Si l'option « Libellés des attributs » est activée, la première cellule de la sélection doit comprendre un en-tête. Si vous n'activez pas cette option, des libellés seront automatiquement créés.

Libellés des sujets : activez cette option si vous voulez utiliser les libellés des sujets pour l'affichage des résultats. Si l'option « Libellés des attributs » est activée, la première cellule de la sélection doit comprendre un en-tête. Si vous n'activez pas cette option, des libellés seront automatiquement créés.

Si le format est **vertical** :

Produits : sélectionnez les produits correspondants aux lignes des données CATA. Si l'option « Libellés des attributs » est activée, la première cellule de la sélection doit comprendre un en-tête.

Sujets : sélectionnez les sujets correspondants aux lignes des données CATA. Si l'option « Libellés des attributs » est activée, la première cellule de la sélection doit comprendre un en-tête.

Sessions : sélectionnez les sessions correspondants aux lignes des données CATA. Si l'option « Libellés des attributs » est activée, la première cellule de la sélection doit comprendre un en-tête.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des attributs : activez cette option si la première ligne des données sélectionnées (Données CATA (0/1), libellés des produits, libellés des sujets) contient un libellé. Si vous n'activez pas cette option, des libellés seront automatiquement créés.

Onglet **Options** :

Filtrer les facteurs : vous pouvez activer l'une des deux options suivantes afin de réduire le nombre de facteurs pour lesquels les résultats sont affichés :

- **% minimum** : activez cette option puis saisissez le pourcentage minimum de la variabilité totale que doivent représenter les facteurs retenus.
- **Nombre maximum** : activez cette option pour fixer le nombre maximum de facteurs à prendre en compte.

Onglet **Données manquantes** :

Ne pas accepter les valeurs manquantes : activez cette option pour que XLSTAT empêche la poursuite des calculs si des valeurs manquantes sont détectées.

Remplacer les valeurs manquantes par 0 : activez cette option si vous considérez que les valeurs manquantes sont équivalentes à des 0.

Onglet **Sorties** :

Statistiques descriptives : activez cette option pour afficher le nombre de cochages par sujet.

Valeurs propres de l'AFC : activez cette option pour afficher le tableau des valeurs propres de l'AFC sur le consensus.

Coordonnées de l'AFC : activez cette option pour afficher les coordonnées du consensus dans l'espace des facteurs.

Matrice de similarité (S) : activez cette option pour afficher la matrice des indices de similarité (Ochiai).

Facteurs de mise à l'échelle : activez cette option pour afficher les facteurs de mise à l'échelle des sujets.

Poids : activez cette option pour afficher les poids créés et utilisés par CATATIS.

Tests des poids : activez cette option pour tester si les poids des sujets sont significatifs.

Configuration consensus : activez cette option pour afficher la configuration consensus créée par CATATIS.

Homogénéité : activez cette option pour afficher l'homogénéité des sujets.

Tests de consistance : activez cette option pour tester si le panel est consistant globalement et par attribut.

Similarité sujets/consensus : activez cette option pour afficher le coefficient de similarité entre chaque sujet et le consensus.

Erreur globale : activez cette option pour afficher l'erreur du critère CATATIS.

Résidu par sujet : activez cette option pour afficher les résidus de CATATIS pour chaque sujet.

Résidu par produit : activez cette option pour afficher les résidus de CATATIS pour chaque produit.

Onglet **Graphiques** :

Valeurs propres de l'AFC : activez cette option pour afficher le graphique (*scree plot*) des valeurs propres de l'AFC sur le consensus.

Biplot de l'AFC : activez cette option pour afficher le graphique des coordonnées du consensus dans l'espace des facteurs.

Graphiques sur 2 axes : activez cette option pour que XLSTAT ne vous demande pas de sélectionner les axes, et affiche automatiquement les 2 premiers.

Facteurs de mise à l'échelle : activez cette option pour afficher le diagramme en bâtons des facteurs de mise à l'échelle des sujets.

Poids : activez cette option pour afficher le diagramme en bâtons des poids créés et utilisés par CATATIS.

Similarité sujets/consensus : activez cette option pour afficher le diagramme en bâtons des coefficients de similarité entre chaque sujet et le consensus.

Résidu par sujet : activez cette option pour afficher le diagramme en bâtons des résidus du critère CATATIS pour chaque sujet.

Résidu par produit : activez cette option pour afficher le diagramme en bâtons des résidus du critère CATATIS pour chaque produit.

Résultats

Répétabilité des sujets : Le coefficient de similarité (Cosinus de Salton) entre les résultats des différentes sessions est affiché. Ce coefficient prend des valeurs entre 0 et 1 et croît avec la ressemblance entre les sessions.

Statistiques descriptives : le nombre de cochages par sujet est affiché. Attention, si vous avez rentré des données non binaires, ce nombre peut être décimal.

Valeurs propres de l'AFC : les valeurs propres de l'AFC et le graphique (*scree plot* correspondant sont affichés.

Coordonnées des produits : les coordonnées des produits du consensus dans l'espace des facteurs sont affichées, ainsi que les graphiques correspondants (en fonction du nombre de facteurs choisi).

Coordonnées des attributs : les coordonnées des attributs du consensus dans l'espace des facteurs sont affichées, ainsi que les graphiques correspondants (en fonction du nombre de facteurs choisi).

Matrice de similarité (S) : la matrice des coefficients de similarité entre tous les sujets est affichée. Le coefficient de similarité utilisé est celui d'Ochiai. Il est compris entre 0 et 1. Plus il est proche de 1, plus la similarité est forte. Cette matrice est utilisée par CATATIS pour calculer les poids des sujets.

Facteurs d'échelle pour chaque sujet : les facteurs d'échelle sont affichés, ainsi que le diagramme en bâtons associé. Plus un facteur d'échelle d'un sujet est grand, plus le nombre de cochages du sujet est faible.

Poids de chaque sujet : les poids calculés par CATATIS sont affichés, ainsi que le diagramme en bâtons associé. Plus un poids est grand, plus le sujet a contribué à l'élaboration du consensus. Sachant que CATATIS donne du poids aux sujets les plus proches du point de vue global, un poids beaucoup plus faible que les autres signifiera que le sujet est atypique.

Tests des poids : les résultats des tests des poids sont affichés. Si un sujet a son poids non significatif, alors son point de vue est très différent du point de vue global, et ses résultats peuvent être remis en cause s'il s'agit d'un expert.

Configuration consensus : la configuration consensus créée par CATATIS est affichée. Elle correspond à la moyenne pondérée des données initiales par les poids de CATATIS.

Homogénéité : l'homogénéité des sujets est affichée. C'est une valeur comprise entre $1/m$ (m étant le nombre de sujets) et 1, qui croît avec l'homogénéité des sujets.

Tests de consistance : les résultats des tests de consistance sont affichés globalement et par attribut. Si le panel est globalement non consistant, les données peuvent malheureusement être jetées. s'il est non consistant pour un ou plusieurs attributs, alors ces attributs sont sujets à tellement de divergence qu'ils ont sûrement été mal compris.

Distance entre la médiane des permutations et l'homogénéité : cette distance permet de voir à quel point l'homogénéité des sujets est forte par rapport à des réponses aléatoires.

Similarité entre chaque sujet et le consensus : les coefficients de similarité entre les sujets et le consensus sont affichés, ainsi que le diagramme en bâtons associé. Tout comme les poids de CATATIS, ces coefficients permettent de détecter des sujets atypiques. L'avantage de ces coefficients est qu'ils sont compris entre 0 et 1, donc plus faciles à interpréter que les poids.

Erreur globale : l'erreur du critère CATATIS est affichée. Elle correspond à la somme de tous les résidus (qui peuvent être présentés par sujet ou par produit).

Résidu par sujet : ce tableau et le diagramme en bâtons correspondant permettent de visualiser la répartition des résidus de CATATIS par sujet. On peut ainsi repérer pour quels sujets CATATIS a été moins efficace, autrement dit, quels sujets se démarquent le plus de la configuration consensus.

Résidu par produit : ce tableau et le diagramme en bâtons correspondant permettent de visualiser la répartition des résidus de CATATIS par produit. On peut ainsi repérer pour quels produits CATATIS a été moins efficace, autrement dit, quels produits se démarquent le plus de la configuration consensus.

Exemple

Un exemple d'utilisation de CATATIS est disponible sur le Centre d'aide XLSTAT à l'adresse suivante :

http://www.xlstat.com/demo-cttf.htm

Bibliographie

Bonnet, L., Ferney, T., Riedel, T., Qannari, E.M., Llobell, F. (September 14, 2022). Using CATA for sensory profiling: assessment of the panel performance. Eurosense, Turku, Finland.

Llobell, F., Cariou, V., Vigneau, E., Labenne, A., Qannari, E. M. (2019). A new approach for the analysis of data and the clustering of subjects in a CATA experiment. *Food Quality and Preference*, **72**, 31-39.

Llobell, F., Giacalone, D., Labenne, A., & Qannari, E. M. (2019). Assessment of the agreement and cluster analysis of the respondents in a CATA experiment. *Food Quality and Preference*, **77**, 184-190.

Llobell, F. (2020). Classification de tableaux de données, applications en analyse sensorielle (Doctoral dissertation, Nantes, Ecole nationale vétérinaire).

CLUSCATA

Utilisez CLUSCATA pour constituer des classes homogènes de sujets sur la base de leurs perceptions des produits.

Dans cette section :

Description

Les épreuves CATA sont très utilisées de nos jours. Cependant, il arrive fréquemment que les perceptions des produits soient différentes entre les sujets. Une classification de ces derniers peut ainsi s'avérer nécessaire. La méthode CLUSCATA s'inscrit dans ce contexte. De plus, cette stratégie permet de mettre de côté les sujets qui ne se conforment à aucune des classes construites. CLUSCATA peut être vue comme une adaptation de CLUSTATIS au cas des données CATA.

Principe de CLUSCATA

L'objectif de CLUSCATA est de constituer des classes de sujets les plus homogènes possible, chaque groupe de sujets étant représenté par un tableau latent (nommée consensus) déterminé par CATATIS. Il est donc naturel que chaque classe soit finalement analysée par CATATIS, afin de déterminer les différences entre les classes constituées. CLUSCATA consiste en un algorithme hiérarchique pouvant être « consolidé » par un algorithme de partitionnement (c'est à dire que l'algorithme de partitionnement est initialisé par la coupe du dendrogramme). Une option intéressante est la création d'une classe « K+1 » (correspondant à une classe supplémentaire) afin de mettre de côté les sujets ne se conformant à aucune classe. Un sujet sera placé dans cette classe si les similarités (coefficents d'Ochiai) entre le consensus de chaque classe et ce sujet sont tous considérés comme faibles.

Structure des données

Il existe deux formats différents :

1. Toutes les données sont concaténées horizontalement (format horizontal).
2. Toutes les données sont concaténées verticalement (format vertical).

Pour la saisie des données, XLSTAT vous demande de sélectionner l'ensemble des données, et de donner le type de format. Dans le cas du format vertical, les produits et les sujets sont demandés.

Interprétation des résultats

La représentation des produits et attributs dans l'espace des k facteurs permet d'interpréter visuellement les proximités entre les produits et les attributs, moyennant certaines précautions.

On peut considérer que la projection d'un produit ou d'un attribut sur un plan est fiable si elle est éloignée du centre du graphique.

Nombre de facteurs

Deux méthodes sont communément utilisées pour déterminer le nombre de facteurs à retenir pour l'interprétation des résultats :

- Regarder la courbe décroissante des valeurs propres. Le nombre de facteurs à retenir correspond au premier point d'inflexion sur la courbe.
- On peut aussi se baser sur le pourcentage cumulé de variabilité représenté par les axes factoriels et décider de se contenter d'un certain pourcentage.

Représentations graphiques

Les représentations graphiques ne sont fiables que si la somme des pourcentages de variabilité associé aux axes de l'espace de représentation est suffisamment élevée. Si ce pourcentage est élevé (par exemple 80%), on peut considérer que la représentation est fiable. Si le pourcentage est faible, il est conseillé de faire des représentations sur plusieurs paires d'axes afin de valider l'interprétation faite sur les deux premiers axes factoriels.

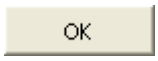
Qualité de la classification

Afin de déterminer la qualité d'une classification hiérarchique, on peut s'aider de l'augmentation de la variance intra-classes (erreur du critère CLUSCATA) provoquée par la fusion de deux classes. Cette augmentation est égale à la hauteur du dendrogramme à laquelle les deux classes de sujets se retrouvent rassemblées dans la même classe.

L'homogénéité de chaque classe et l'homogénéité globale sont également des indices très importants (entre $1/m$ et 1, m étant le nombre de sujets) qui permettent de juger de la qualité de la classification. Il est à noter que la consolidation et l'ajout d'une classe « K+1 » peuvent augmenter les homogénéités.

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.


: cliquez sur ce bouton pour changer la façon dont vous chargez les données dans XLSTAT. Si le bouton possède une icône de souris, XLSTAT vous permet de faire une sélection à la souris des données. Si le bouton possède une icône de liste, XLSTAT vous permet de sélectionner les données à partir d'une liste. Si le bouton possède une icône orange, des boutons complémentaires avec un point d'interrogation  sont affichés pour vous permettre d'importer les données à partir de fichiers.

Onglet **Général** :

Données CATA (0/1) : sélectionnez les données correspondant aux différents sujets. Si la première ligne de la sélection comprend des en-têtes, l'option « Libellés des attributs » doit être activée.

Format : Cliquez sur horizontal ou vertical selon la façon dont sont structurées vos données.

Si le format est **horizontal** :

Nombre de sujets : entrez le nombre de sujets dans les données CATA (format horizontal uniquement).

Libellés des produits : activez cette option si vous voulez utiliser des libellés des produits pour l'affichage des résultats. Si l'option « Libellés des attributs » est activée, la première cellule de la sélection doit comprendre un en-tête. Si vous n'activez pas cette option, des libellés seront automatiquement créés.

Libellés des sujets : activez cette option si vous voulez utiliser les libellés des sujets pour l'affichage des résultats. Si l'option « Libellés des attributs » est activée, la première cellule de la sélection doit comprendre un en-tête. Si vous n'activez pas cette option, des libellés seront automatiquement créés.

Si le format est **vertical** :

Produits : sélectionnez les produits correspondants aux lignes des données CATA. Si l'option « Libellés des attributs » est activée, la première cellule de la sélection doit comprendre un en-tête.

Sujets : sélectionnez les sujets correspondants aux lignes des données CATA. Si l'option « Libellés des attributs » est activée, la première cellule de la sélection doit comprendre un en-tête.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des attributs : activez cette option si la première ligne des données sélectionnées (Données CATA (0/1), libellés des produits, libellés des sujets) contient un libellé. Si vous n'activez pas cette option, des libellés seront automatiquement créés.

Onglet **Options** :

Filtrer les facteurs : vous pouvez activer l'une des deux options suivantes afin de réduire le nombre de facteurs pour lesquels les résultats sont affichés :

- **% minimum** : activez cette option puis saisissez le pourcentage minimum de la variabilité totale que doivent représenter les facteurs retenus.
- **Nombre maximum** : activez cette option pour fixer le nombre maximum de facteurs à prendre en compte.

Troncature : activez cette option si vous voulez que XLSTAT définisse **automatiquement** une troncature, et donc le nombre de classes à retenir, ou si vous voulez définir vous-même le **nombre de classes** à créer, ou si vous voulez définir le **niveau** auquel le dendrogramme doit être tronqué.

Consolidation : activez cette option pour réaliser une consolidation des classes obtenues à partir du dendrogramme.

Classe K+1 : activez cette option pour ajouter une classe supplémentaire qui contiendra les sujets ne se conformant à aucune des classes.

Paramètre rho : Ce paramètre représente l'accord minimal pour être considéré comme suffisamment en accord pour être conservé dans une classe. Plus ce paramètre augmente, plus l'accord demandé avec la classe est fort et plus vous risquez de placer de sujets dans la classe K+1. Choisissez la façon dont vous voulez définir le paramètre rho : **automatiquement** ou **défini par l'utilisateur**.

Onglet **Données manquantes** :

Ne pas accepter les valeurs manquantes : activez cette option pour que XLSTAT empêche la poursuite des calculs si des valeurs manquantes sont détectées.

Remplacer les valeurs manquantes par 0 : activez cette option si vous considérez que les valeurs manquantes sont équivalentes à des 0.

Onglet **Sorties** :

Statistiques descriptives : activez cette option pour afficher le nombre de cochages par sujet.

Matrice de similarité (S) : activez cette option pour afficher la matrice des indices de similarité (Ochiai).

Statistiques des nœuds : activez cette option pour afficher les statistiques des nœuds du dendrogramme.

Composition des classes : activez cette option pour afficher la composition de chacune des classes.

Valeurs propres de l'AFC : activez cette option pour afficher le tableau des valeurs propres de l'AFC sur le consensus de chaque classe.

Coordonnées de l'AFC : activez cette option pour afficher les coordonnées de l'AFC de chaque classe dans l'espace des facteurs.

Configuration consensus : activez cette option pour afficher la configuration consensus de chaque classe créée par CATATIS.

Similarité sujets/consensus : activez cette option pour afficher le coefficient de similarité entre chaque sujet et le consensus de sa classe.

Poids : activez cette option pour afficher les poids des sujets créés et utilisés par CATATIS dans chaque classe.

Homogénéités : activez cette option pour afficher l'homogénéité de chacune des classes ainsi que l'homogénéité globale.

Erreur globale/Variance intra-classes : activez cette option pour afficher l'erreur du critère de minimisation CLUSCATA, équivalente à la variance intra-classes.

Onglet **Graphiques** :

Diagramme des niveaux : activez cette option pour afficher le diagramme des niveaux permettant d'observer l'impact des regroupements successifs sur la variance intra-classes.

Dendrogramme : activez cette option pour afficher le dendrogramme.

- **Horizontal** : choisissez cette option pour afficher un dendrogramme horizontal.
- **Vertical** : choisissez cette option pour afficher un dendrogramme vertical.
- **Tronqué** : activez cette option pour afficher le dendrogramme tronqué (le dendrogramme commence au niveau de la troncature).
- **Etiquettes** : activez cette option pour afficher les libellés des sujets (dendrogramme complet) ou des classes (dendrogramme tronqué) sur le dendrogramme.

- **Couleurs** : activez cette option pour utiliser des couleurs pour représenter les différents groupes sur le dendrogramme complet.

Graphiques sur deux axes : activez cette option si vous souhaitez que les différentes représentations graphiques ne soient affichées que sur les deux premiers axes.

Valeurs propres de l'AFC : activez cette option pour afficher le graphique (*scree plot*) des valeurs propres de l'AFC sur le consensus de chaque classe.

Biplot de l'AFC : activez cette option pour afficher le graphique des coordonnées du consensus de chaque classe dans l'espace des facteurs.

Similarité sujets/consensus : activez cette option pour afficher le diagramme en bâtons des coefficients de similarité entre chaque sujet et le consensus de sa classe.

Poids : activez cette option pour afficher le diagramme en bâtons des poids créés et utilisés par CATATIS dans chaque classe.

Résultats

Statistiques descriptives : le nombre de cochages par sujet est affiché.

Matrice de similarité (S) : la matrice des coefficients de similarité entre tous les sujets est affichée. Le coefficient de similarité utilisé est celui d'Ochiai. Il est compris entre 0 et 1. Plus il est proche de 1, plus la similarité est forte. Cet indice est le coefficient d'Ochiai.

Statistiques des nœuds : dans ce tableau sont affichées les informations concernant les nœuds successifs du dendrogramme. Le premier nœud a pour indice le nombre de sujets augmenté de 1. Ainsi, il est aisé de repérer à quel moment un sujet ou un groupe de sujets est regroupé avec un autre groupe de sujets dans le dendrogramme.

Diagramme des niveaux : dans ce graphique sont affichés les niveaux des nœuds du dendrogramme, qui correspondent à l'augmentation du critère de minimisation de CLUSCATA (équivalent à l'augmentation de la variance intra-classes) lors de la fusion de deux classes.

Dendrogrammes : le dendrogramme complet permet de visualiser le regroupement progressif des sujets. Si une troncature a été demandée, un trait en pointillé marque le niveau auquel est effectuée la troncature. Le dendrogramme tronqué permet de visualiser les classes après la troncature.

Compositions des classes :

Résultats par sujet : dans ce tableau est indiquée pour chaque sujet sa classe d'affectation dans l'ordre initial des sujets. Si une consolidation est demandée, les résultats sont donnés avant et après la consolidation. Dans le cas où vous avez coché classe « K+1 », il est possible que certains sujets aient une valeur manquante après la consolidation. Ceci signifie qu'ils ne sont placés dans aucune des classes principales (ils sont placés dans la classe « K+1 »).

Résultats par classe : Les résultats sont donnés par classe. Ainsi, une liste de sujets est affichée pour chacune des classes.

Nombre de sujets par classe : Le nombre de sujets dans chaque classe est indiqué.

Paramètre rho calculé : Résultat affiché uniquement si vous avez choisi d'ajouter une classe « K+1 ». Le paramètre rho représente la similarité minimale que doit avoir un sujet avec le consensus d'une classe pour lui appartenir. Si cette condition n'est respectée pour aucune des classes, le sujet est placé dans la classe « K+1 ». Ce paramètre est calculé en fonction de la proximité de chaque sujet avec sa classe ainsi qu'avec la classe voisine.

Analyse de la classe k :

Dans cette section est affichée l'analyse de chacune des classes par la méthode CATATIS. Chaque classe est analysée tour à tour.

Valeurs propres de l'AFC : les valeurs propres de l'AFC et le graphique (*scree plot*) correspondant sont affichés.

Coordonnées des produits : les coordonnées des produits du consensus dans l'espace des facteurs sont affichées, ainsi que les graphiques correspondants (en fonction du nombre de facteurs choisi).

Coordonnées des attributs : les coordonnées des attributs du consensus dans l'espace des facteurs sont affichées, ainsi que les graphiques correspondants (en fonction du nombre de facteurs choisi).

Configuration consensus : la configuration consensus créée par CATATIS est affichée. Elle correspond à la moyenne des sujets pondérée par les poids de CATATIS.

Similarité entre chaque sujet et le consensus : les coefficients de similarité entre les sujets et le consensus sont affichés, ainsi que le diagramme en bâtons associé. Tout comme les poids de CATATIS, ces coefficients permettent de détecter des sujets atypiques. L'avantage de ces coefficients est qu'ils sont compris entre 0 et 1, donc plus faciles à interpréter que les poids.

Poids de chaque sujet : les poids calculés par CATATIS sont affichés, ainsi que le diagramme en bâtons associé. Plus un poids est grand, plus le sujet a contribué à l'élaboration du consensus. Sachant que CATATIS donne du poids aux sujets les plus proches du point de vue global, un poids beaucoup plus faible que les autres signifiera que le sujet est atypique.

Indices :

Homogénéités : l'homogénéité de chaque classe est affichée. C'est une valeur comprise entre $1/m$ (m étant le nombre de sujets de la classe) et 1, qui croît avec l'homogénéité des sujets. Dans un second temps, l'homogénéité globale, qui est une moyenne pondérée des homogénéités de chaque classe, est affichée.

Erreur globale/Variance intra-classes : l'erreur du critère CLUSCATA est affichée. Elle correspond à la variance intra-classes.

Exemple

Un exemple d'utilisation de CLUSCATA est disponible sur le Centre d'aide XLSTAT à l'adresse suivante :

http://www.xlstat.com/demo-cscf.htm

Bibliographie

Llobell, F., Cariou, V., Vigneau, E., Labenne, A., Qannari, E. M. (2019). A new approach for the analysis of data and the clustering of subjects in a CATA experiment. *Food Quality and Preference*, **72**, 31-39.

Llobell, F., Giacalone, D., Labenne, A., & Qannari, E. M. (2019). Assessment of the agreement and cluster analysis of the respondents in a CATA experiment. *Food Quality and Preference*, **77**, 184-190.

Llobell, F. (2020). Classification de tableaux de données, applications en analyse sensorielle (Doctoral dissertation, Nantes, Ecole nationale vétérinaire).

Graphiques sémantiques différentiels

Utilisez cette méthode pour visualiser les notes attribuées par des sujets à des objets pour différents critères.

Dans cette section :

Description

Le psychologue Charles E. Osgood a développé à la fin des années 1950 une méthode de visualisation dénommée Semantic differential dans le but de représenter graphiquement les différentes connotations associées à un mot par différents individus. Osgood a demandé aux participants de ses études de noter un mot sur une série d'échelles allant d'un extrême à l'autre (par exemple favorable/défavorable). De la distance observée entre les différents profils observés pour des individus ou des groupes d'individus, Osgood a déduit la distance psychologique et éventuellement comportementale entre les individus ou les groupes.

Cette méthode peut aussi être appliquée dans d'autres situations :

- Analyse des perceptions d'experts à propos d'un produit (par exemple un yaourt) décrit par divers attributs (par exemple, acidité, salé, sucré, texture, …) sur des échelles similaires (soit d'un extrême à l'autre, soit sur les échelles de notation). La visualisation en graphique sémantique différentiel permet de rapidement identifier s'il y a des différences entre les experts et à quel niveau se situent les différences.
- Analyse d'enquêtes de satisfaction.
- Analyse des profils de candidats dans le cadre d'un recrutement.

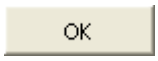
Cet outil peut être utilisé en analyse sensorielle. Voici deux exemples d'application dans ce contexte :

- Un panel de sujets note un produit alimentaire sur une échelle ordinale (codée de 1 à 5) en fonction de plusieurs critères (les « attributs ») tels que la texture, l'apparence visuelle, l'odeur, le goût, le prix etc. Dans ce cas, le tableau de données sera constitué de telle sorte que la case (i,j) du tableau corresponde à la note donnée au produit par le sujet i pour l'attribut j. Le graphique sémantique différentiel permet alors de comparer visuellement les sujets.

- Un panel de sujets note des produits alimentaires (les « objets ») sur une échelle ordinale (codée de 1 à 5) en fonction de plusieurs critères (les « attributs ») tels que la texture, l'apparence visuelle, l'odeur, le goût, le prix etc. Dans ce cas, le tableau de données sera constitué de telle sorte que la case (i,j) du tableau corresponde à la note moyenne donnée par les sujets au produit i pour l'attribut j. Le graphique sémantique différentiel permet de comparer visuellement les produits pour les différents attributs.

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.

Onglet **Général** :

Descripteurs : sélectionnez les données de descripteurs sur la feuille Excel. Si la première ligne de la sélection comprend des en-têtes, l'option « Libellés des descripteurs » doit être activée.

Objets : sélectionnez les libellés des objets sur la feuille Excel.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des descripteurs : activez cette option si la première ligne des données sélectionnées (données et libellés des observations) contient un libellé.

Onglet **Graphiques** :

Couleur : activez cette option pour utiliser une couleur différente pour chacun des objets/individus/experts.

Quadrillage : activez cette option pour afficher le quadrillage sur le graphique.

Valeurs : activez cette option pour indiquer les valeurs sur le graphique.

Résultats

Le résultat affiché est le graphique sémantique différentiel. Comme il s'agit d'un graphique Excel, vous pouvez ensuite modifier à votre guise les différents éléments.

Exemple

Un exemple de graphique sémantique différentiel est disponible sur le Centre d'aide XLSTAT :

http://www.xlstat.com/demo-sdf.htm

Bibliographie

Judd C.M., Smith E.R. and Kidder L.H (1991). Research Methods in Social Relations. Holt, Rinehart & Winston, New York.

Osgood C.E., Suci G.J. and Tannenbaum P.H. (1957). The Measurement of Meaning. University of Illinois Press, Urbana.

Oskamp S. (1977). Attitudes and Opinions. Prentice-Hall, Englewood Cliffs, New Jersey.

Snider J. G. and Osgood C.E. (1969). Semantic Differential Technique. A Sourcebook. Aldine Press, Chicago.

Analyse TURF

Utilisez cet outil pour effectuer une analyse TURF (Total Unduplicated Reach and Frequency) pour un ensemble de produits, et mettre en avant une ligne de produits qui aura une meilleure pénétration du marché.

Dans cette section :

Description

XLSTAT permet d'appliquer la méthode TURF (Total Unduplicated Reach and Frequency) qui est largement utilisée en marketing. Cette méthode sert généralement à mettre en avant un sous-ensemble de produits issus d'une gamme complète, qui permettra une pénétration la plus complète possible du marché. Ainsi, sur l'ensemble des produits, on pourra obtenir un sous-ensemble, qui pourra constituer une ligne de produits pour laquelle on pourra espérer obtenir le plus de parts de marché possible (qui touchera le plus de consommateurs possible, d'où le terme « reach »).

Par exemple, l'utilisation de cette méthode peut s'appliquer chez un fabriquant de glaces. Le fabriquant produit 30 parfums mais il désire mettre en avant les 6 qui plairont au plus de clients possible. Ainsi, il soumet un questionnaire à un panel de 500 consommateurs qui notent chaque parfum sur une échelle de 1 à 10. On considère que le consommateur sera satisfait et enclin à consommer le parfum si il donne une note au-delà de 8. Ainsi, il va lancer avec XLSTAT une analyse TURF afin de recherche la combinaison de 6 parfums pour lesquels il y a au moins une fois une note au-delà de 8 chez les consommateurs. Il obtiendra donc un groupe de parfums qui plairont au plus grand nombre.

En se basant sur un questionnaire (avec des notes sur une échelle fixe), on recherche les combinaisons de produits permettant de toucher le plus de personnes possibles.

XLSTAT propose différentes techniques afin de trouver la meilleure combinaison possible :

- l'énumération qui va tester l'ensemble des combinaisons mais qui peut s'avérer très coûteuse en terme de temps de calcul,
- l'algorithme glouton qui est très rapide mais qui n'assure pas de trouver l'optimum
- un algorithme rapide qui recherche aussi la meilleure combinaison mais qui ne garantit pas la solution optimale.

Méthodes :

Les données utilisées doivent être des données issues d'un questionnaire : une ligne par consommateur et une colonne par produit. Elles doivent être sous forme de notes (échelles de Likert). XLSTAT permet de définir différentes échelles. Néanmoins, toutes les notes doivent être sur la même échelle. L'utilisateur choisit un intervalle pour lequel on considère que l'objectif est atteint (par exemple les notes plus grandes que 8 sur 10).

XLSTAT permet d'utiliser 3 algorithmes différents afin de trouver la bonne ligne de produits :

- **L'énumération**

On va tester toutes les combinaisons de k produits parmi les p produits présents ($k < p$). On conservera les combinaisons qui ont le « reach » le plus élevé. Le « reach » est défini par :

Reach = Nombre de fois pour lesquelles l'objectif est atteint au moins une fois pour un consommateur pour la combinaison analysée.

Cette méthode est exacte mais peut être très longue lorsque les nombres p et k deviennent grands (par exemple pour p=40 produits et k=12 produits dans le sous-ensemble, on aura 5 586 853 480 combinaisons).

- **L'algorithme glouton**

Cet algorithme est une heuristique simple qui permet de trouver un bon résultat très rapidement en maximisant le reach.

Il fonctionne de la manière suivante :

- Trouver le produit qui atteint l'objectif le plus souvent

- Sélectionner ce produit dans la combinaison

- Retirer les observations pour lesquelles ce produit a atteint l'objectif

- Répéter jusqu'à ne plus avoir d'observations à retirer ou jusqu'à ne plus pouvoir en retirer

Cet algorithme est répété de nombreuses fois avec des conditions initiales différentes afin de minimiser le risque de tomber dans des optimums locaux. Son avantage réside dans sa vitesse mais il ne garantit pas d'obtenir le maximum global.

- **L'algorithme accéléré**

Cet algorithme part du même principe que l'énumération, mais lorsqu'aucune amélioration n'est trouvé au bout d'un certain nombre de combinaisons, un saut dans les combinaisons permettant d'éviter un certain nombre de combinaisons « inutiles » est effectué. Cet algorithme ne garantit pas non plus de trouver l'optimum local mais il explorera plus de possibilités que l'algorithme glouton.

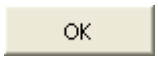
Les contraintes pour les analyses TURF

XLSTAT propose d'imposer des contraintes dans le cadre des analyses TURF. Deux types de contraintes sont disponibles :

- Contraintes d'appartenance : dans ce cas, on force l'appartenance d'un produit à la ligne de produit.
- Contraintes de présence d'éléments appartenant à un groupe : dans ce cas, on aura une variable supplémentaire associant un groupe à chaque produit. Lors de la recherche de solution, la ligne obtenue aura au moins un produit de chaque groupe.

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.

Onglet **Général**:

Données : sélectionnez les données sous forme de notes sur une échelle commune pour tous les produits. Si des en-têtes ont été sélectionnés, veuillez vérifier que l'option « Libellés des échantillons » est activée.

Taille du sous-ensemble : sélectionnez la taille du sous-ensemble (c'est- à-dire le nombre de produits qui doivent être intégrés dans la ligne de produits).

Autre groupe : activez cette option si vous désirez sélectionner un autre groupe de produits dans lequel un sous-ensemble sera aussi obtenu. Il arrive que sur l'ensemble des produits d'une marque on ait deux catégories de produits. On voudra un certain nombre de produit de

chacune des catégories pour former sa ligne de produits. (Dans le cas de l'exemple précédent, on pourra vouloir 3 sorbets et 3 glaces.)

Echelle : choisissez l'échelle utilisée pour noter les produits. Si vous sélectionnez l'option « autres », vous devrez alors entrer le minimum et le maximum de votre échelle.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des variables : activez cette option si la première ligne des sélections (données, autre groupe) contient un libellé.

Libellés des observations : activez cette option si vous voulez utiliser des libellés d'observations disponibles sur une feuille Excel pour l'affichage des résultats. Si l'option « Libellés des variables » est activée, la première cellule de la sélection doit comprendre un en-tête.

Poids des observations : activez cette option si vous voulez utiliser des poids associé aux observations disponibles sur une feuille Excel pour l'affichage des résultats. Si l'option « Libellés des variables » est activée, la première cellule de la sélection doit comprendre un en-tête.

Contraintes sur les produits : activez cette option si vous voulez que certains produits soient automatiquement ajouter dans les lignes de produits générées par l'analyse TURF. Si cette option est activée, une fois que vous avez cliqué sur Ok, une fenêtre vous permettant de choisir les produits à ajouter apparaît.

Contraintes de groupes : activez cette option si vous voulez qu'au moins un produit de chaque groupe soit sélectionné dans chaque ligne de produit. Sélectionnez la colonne dans laquelle les groupes associés aux produits sont affichés. Cette colonne doit avoir autant d'éléments que de produit. Si des en-têtes ont été sélectionnés, veuillez vérifier que l'option « Libellés des échantillons » est activée.

Onglet **Options**:

Nombre de combinaisons affichées : entrez le nombre de combinaisons que vous voulez conserver. Il est parfois intéressant de regarder plusieurs combinaisons qui obtiennent de bon « reach » afin de sélectionner la meilleure ligne de produits.

L'objectif est atteint pour des scores entre __ et __ : entrez les bornes inférieures et supérieures des scores qui permettront de considérer que l'objectif (« reach ») est atteint.

Méthode : sélectionnez la méthode que vous voulez utiliser pour l'analyse TURF. Pour l'énumération, vous pouvez entrer un temps maximum pour que l'algorithme s'arrête

automatiquement. Si le nombre de combinaisons est réduit, XLSTAT opte automatiquement pour la méthode d'énumération car celle-ci est exacte.

Onglet **Données manquantes** :

Ne pas accepter les valeurs manquantes : activez cette option pour que XLSTAT empêche la poursuite des calculs si des valeurs manquantes sont détectées.

Supprimer les observations : activez cette option pour supprimer les observations comportant des données manquantes.

Résultats

Fréquences par produit : dans ce tableau sont affichées les fréquences auxquelles l'objectif a été atteint pour chaque produit.

Lignes de produits obtenus par l'analyse TURF : dans ce tableau sont affichées pour chaque combinaison sélectionnée : le reach, la fréquence et le nom de chacun des produits conservés.

Lignes de produits obtenus par l'analyse TURF (%) : dans ce tableau sont affichées pour chaque combinaison sélectionnée : le pourcentage d'observation pour lesquelles l'objectif a été atteint, la fréquence en pourcentage, et la fréquence en pourcentage pour chaque produit dans chacune des combinaisons.

Tableau croisé (TURF) : ce tableau est un tableau récapitulatif de l'analyse avec les lignes obtenues par colonnes et les produits par lignes. La première ligne du tableau donne la valeur du reach pour chaque ligne de produits et la seconde le reach en pourcentage. La dernière colonne donne la fréquence d'apparition en pourcentage d'un produit dans toutes les lignes. Finalement, lorsqu'un produit est présent dans une ligne de produits, la valeur affichée est égal à la fréquence du produit dans les données divisé par le reach de la ligne de produits analysé (multiplié par 100).

Aucune intention d'achat : ce résultat permet de savoir le nombre de sujet(s) n'ayant émis aucune intetion d'achat.

Exemple

Un exemple d'analyse TURF est disponible sur le Centre d'aide XLSTAT :

http://www.xlstat.com/demo-turff.htm

Bibliographie

Miaoulis G., V. Free and H. Parsons (1990). TURF: A New Planning Approach For Product Line Extensions, *Marketing Research*, II (March), 28-40.

Krieger A. M. and P. E. Green (2000). TURF Revisited: Enhancements to Total Unduplicated Reach and Frequency Analysis, *Marketing Research*, 12 (Winter), 30-36.

Roue sensorielle

Utilisez l'outil roue sensorielle pour représenter sur un diagramme de type anneaux imbriqués une classification de termes décrivant un produit.

Dans cette section :

Description

L'outil roue sensorielle permet de représenter sur un diagramme synthétique de type anneaux imbriqués, une classification de termes utilisés pour décrire un produit. Une représentation alternative pour ce type de données est un arbre. Néanmoins l'idée de bifurcation des arbres induit une idée d'alternative, ce qui n'est pas forcément le cas dans la description d'un produit. Par exemple, dans une roue sensorielle, on peut décider que la partie gauche concerne le goût et la partie droite la vue, sans qu'une idée d'alternative vienne polluer la compréhension : la partie gauche comprendra les termes concernant le goût et, celle de droite, les mots décrivant la perception visuelle.

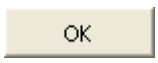
La classification doit idéalement être conçue de telle sorte qu'un terme ne soit présent qu'à un seul endroit sur le diagramme.

XLSTAT permet de partir de deux formats de données alternatifs :

- une table de classification décrivant une hiérarchie entre les termes ;
- une liste de mots uniques, avec éventuellement des fréquences ou des poids associés à chaque mot (par défaut les poids relatifs sont considérés comme identiques). Si la liste comprend des termes répétés, on peut alors prendre en compte les répétitions ou non.

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

Annuler : cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

Aide : cliquez sur ce bouton pour afficher l'aide.

 : cliquez sur ce bouton pour rétablir les options par défaut.

 : cliquez sur ce bouton pour effacer les sélections de données.

 : cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.

Onglet **Général**:

Données : sélectionnez les données en respectant l'un des deux **formats** proposés : vous pouvez soit sélectionner une **liste de mots**, soit sélectionner un **tableau de classification**. Si des en-têtes ont été sélectionnés, veuillez vérifier que l'option « Libellés des échantillons » est activée.

Effectifs : activez cette option si vous voulez affecter aux mots des effectifs ou des poids non égaux.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés inclus : activez cette option si la première ligne des sélections (données et effectifs) contient un libellé.

Onglet **Options**:

Reformater les mots : activez cette option pour que XLSTAT, d'une part identifie les mots répétés et, d'autre part, élimine les espaces avant et après chaque mot.

Dimensionner en fonction de la taille : activez cette option pour que XLSTAT dimensionne les secteurs de la roue sensorielle en fonction de la fréquence des mots.

Nombre de niveaux : entrez le nombre de niveaux à afficher sur la roue sensorielle.

Onglet **Données manquantes** :

Ne pas accepter les valeurs manquantes : activez cette option pour que XLSTAT empêche la poursuite des calculs si des valeurs manquantes sont détectées.

Supprimer les observations : activez cette option pour supprimer les observations comportant des données manquantes.

Onglet **Sorties** :

Fréquence des mots : activez cette option pour que XLSTAT affiche avant la roue sensorielle la fréquence des mots (en %).

Résultats

Si l'option **fréquence des termes** a été activée, dans le premier tableau sont affichés les différents mots et les effectifs et % associés.

La roue sensorielle est ensuite affichée. Le bouton qui la précède permet d'activer ou désactiver le graphique. Lorsqu'il est activé, un simple clic sur le graphique permet d'afficher une boîte de dialogue qui permet cinq actions différentes :

- Fusionner deux termes
- Déplacer un terme après un autre
- Renommer un terme
- Aligner l'orientation des termes sur les rayons
- Aligner l'orientation des termes perpendiculairement aux rayons

Exemple

Un exemple de création de roue sensorielle est disponible sur le Centre d'aide XLSTAT :

http://www.xlstat.com/demo-sensowheelf.htm

Bibliographie

Meilgaard M.C., Da Igliesh C.E. and J.F. Clapperton (1979). Progress towards an international system of beer flavour terminology. *Journal of American Society of Brewing Chemists* **, 37**, 42-52.

Piggot J.R. and Jardine S.P. (1979). Descriptive sensory analusis of whiskey flavour. *The Journal of the Institute of Brewing and Distilling*, **85**, 82-85.

Plans d'expériences pour l'analyse sensorielle

Utilisez cet outil pour créer un plan d'expériences optimal, ou quasi-optimal, dans le cadre d'expériences visant à modéliser les préférences d'un ensemble de consommateurs ou d'experts pour différents produits.

Dans cette section :

Description

La planification expérimentale est une étape fondamentale pour quiconque veut s'assurer que les données collectées seront exploitables dans les meilleures conditions statistiques possibles. Rien ne sert de faire évaluer des produits par un panel de sujets si l'on ne peut ensuite comparer les produits dans des conditions statistiques satisfaisantes. Il n'est par ailleurs pas nécessaire de faire évaluer tous les produits par tous les sujets pour pouvoir comparer les produits entre eux.

Cet outil a pour but de permettre aux spécialistes de l'analyse sensorielle de disposer d'un outil simple et puissant pour mettre en place une étude sensorielle menée auprès de sujets (experts et/ou consommateurs) évaluant un ensemble de produits.

Lorsque l'on veut faire évaluer par exemple 9 produits par un panel de consommateurs, se pose dans un premier temps la question du nombre de consommateurs à faire intervenir, sachant qu'il peut exister des contraintes techniques (on n'a accès qu'à un nombre limité de consommateurs entrainés) ou des contraintes budgétaires. Une fois le nombre de consommateurs arrêté, par exemple 82, se pose la question du nombre maximum de produits que peut évaluer un consommateur à chaque session, rarement pour des raisons budgétaires, mais plus souvent en raison de contraintes physiologiques : un consommateur, même entrainé, ne peut pas nécessairement garder toutes ses capacités sensorielles s'il évalue trop de produits à la suite. Imaginons que l'expérience montre que trois produits est un maximum pour une session et que pour des raisons d'organisation, seules deux sessions peuvent être organisées. Chaque consommateur évaluera donc au maximum 6 produits.

Reste maintenant à déterminer quels produits seront évalués par chacun des 82 consommateurs, au cours de chaque session, et dans quel ordre. Il est possible que l'ordre ait une influence (ce n'est pas le sujet ici, mais un plan d'expérience pourrait permettre dans une autre étude de vérifier ou invalider cette hypothèse). Afin d'éviter que certains produits soient pénalisés, il faut donc de faire en sorte que les produits soient vus aussi souvent que possible dans les 3 positions possibles au cours de chaque session. Par ailleurs, il est possible que

certains enchainements de produits aient aussi une influence sur les appréciations sensorielles. On se limite ici à considérer les couples ordonnés de produits (carry-over d'ordre 2). Comme pour l'ordre, on veillera aussi à ce que les différents couples ordonnés, 72 dans notre exemple, soient présents avec une fréquence aussi homogène que possible dans le plan d'expériences.

La génération du plan va donc essayer de concilier la triple exigence suivante :

- Les produits doivent être vus par autant de sujets que possible et avec une fréquence globale pour les différents produits aussi homogène que possible,
- Chaque produit doit être vu dans les différentes positions au cours de chaque session, avec une fréquence globale pour chaque couple (position, produit) aussi homogène que possible
- Les différents couples ordonnés de produits doivent être présents dans le plan d'expériences avec une fréquence aussi homogène que possible.

Mesure de la performance du plan

Appelons N la matrice ayant autant de lignes que de produits et autant de colonnes que de sujets, et contenant le nombre de fois où chaque sujet voit chaque produit au cours de chaque session. Dans les plans sensoriels que nous traitons ici, N contient soit des 0, soit des 1 : on impose qu'un sujet voit chaque produit au maximum une fois. Par ailleurs les sommes marginales par colonnes sont constantes et égales à k (le nombre de produits vus par chaque sujet est imposé).

La matrice $M = NN'$ a la particularité de comporter sur sa diagonale la fréquence de chaque produit dans le plan, et sur les parties triangulaires inférieures et supérieures le nombre de fois où des sujets ont évalué chaque couple de produits (ici l'ordre n'importe pas). On l'appelle **matrice de cooccurrence** (*concurrence matrix* en anglais).

Soit la matrice $A^* = I - \frac{qNN'q}{k}$ où q est une matrice diagonale comportant l'inverse de la racine carrée de la fréquence de chaque produit. On montre que cette matrice est directement liée à la matrice d'information concernant les produits et donc aux variances et covariances des paramètres associés aux produits dans un modèle d'ANOVA que l'on pourra calculer une fois les évaluations recueillies. Si l'on veut s'assurer que les variances des différences entre les paramètres associés aux produits soient aussi homogènes que possible, on montre qu'il faut faire en sorte que les valeurs propres de la matrice A^* soit aussi proches les unes des autres.

On définit la **A-efficacité** comme la moyenne harmonique des, au plus p-1, valeurs propres non nulles de la matrice A^*, et la **D-efficacité** comme la moyenne géométrique des mêmes valeurs propres. Les deux critères sont égaux dans le cas idéal où toutes les valeurs propres sont égales.

Plans en blocs incomplets équilibrés

Un plan en blocs est un plan d'expériences dans lequel on étudie l'influence d'au moins deux facteurs sur un ou plusieurs phénomènes. On sait que l'un des facteurs a par construction un

effet important, sans que l'on puisse agir dessus, mais ce n'est pas celui qui nous intéresse. On veut donc pouvoir s'assurer que ce facteur ne perturbera pas les analyses que l'on effectuera une fois les données collectées. Pour cela on fait en sorte que les différents niveaux des autres facteurs soient aussi bien représentés dans chacun des blocs (les modalités du facteur bloc).

Dans notre cas, nous avons un facteur bloc qui correspond aux sujets, et un facteur que l'on souhaite particulièrement étudié, le facteur produit.

Un plan en blocs complets est un plan dans lequel tous les niveaux des facteurs étudiés sont présents une fois à l'intérieur de chaque bloc. Cela correspond, pour un plan sensoriel, au cas où tous les produits sont vus une fois par l'ensemble des sujets.

Un plan en blocs incomplets est un plan dans lequel tous les niveaux des facteurs étudiés ne sont pas présents dans chaque bloc. Il est équilibré si chaque niveau de chaque facteur étudié est présent un même nombre r de fois et si chaque couple de niveaux de chaque facteur étudié est présent un même nombre de fois λ.

Si v est le nombre de produits étudiés, b le nombre de sujets, k le nombre de produits vus par chaque sujet, on montre que les conditions suivantes sont nécessaires (mais non suffisantes) pour avoir un plan en blocs incomplets équilibrés :

$$bk = vr$$

$$r(k-1) = \lambda(v-1)$$

Dans le cas où un plan en blocs incomplets équilibrés existe, on connaît la valeur optimale des deux critères (A et D-efficacité). Cette valeur est donnée par

$$E = \frac{\nu(k-1)}{k(\nu-1)}$$

XLSTAT permet de chercher un plan optimal au sens de la A- efficacité ou de la D- efficacité tant dans le cas des plans complets que dans le cas des plans en blocs incomplets, qu'ils soient équilibrés ou non.

Algorithmes de recherche du plan

XLSTAT s'appuie sur deux techniques différentes pour générer les plans. Si l'option rapide est choisie par l'utilisateur et si $\left(\frac{b}{v}\right)$ est entier, alors XLSTAT utilise la méthode des plans cycliques introduite par Williams (1949), et étudiée très en détail par John et Williams (1995). Si (b/v) n'est pas entier ou si l'option de calcul rapide n'est pas demandée, XLSTAT utilise une méthode propriétaire (non publiée) très performante permettant de générer très vite une solution pertinente et à partir de cette solution, de chercher par recuit simulé une meilleure solution, pendant un temps maximal fixé par l'utilisateur. Si le plan recherché est un plan en blocs complets ou incomplets équilibrés et que l'optimum est trouvé, la recherche de plan s'interrompt avant que le temps imparti soit écoulé.

Algorithmes d'amélioration des fréquences des positions (effet d'ordre ou de position) et du carry-over (effet de report)

Une fois le plan trouvé (la matrice N est dorénavant connue), reste à ordonner les produits de façon à optimiser le plan sensoriel (Périnel et Pagès, 2004). On veut notamment que chaque produit se trouve un nombre de fois égal à chaque rang et que chaque couple ordonné de produits se retrouve un nombre égal de fois. XLSTAT va se servir pour cela de deux matrices : la matrice des fréquences des positions et la matrice du carry-over. Le but de l'algorithme sera de rendre le plus homogène possible ces matrices. La matrice de carry- over est une matrice qui fait apparaître à la position ij le nombre de fois que le produit i précède le produit j dans le plan. On définit un paramètre λ qui va nous permettre de favoriser soit l'obtention d'un bon carry-over (λ proche de 0) soit l'obtention d'une matrice des fréquences de positions le plus proche de la matrice constante (λ proche de 1).

L'algorithme d'optimisation est itératif et consiste en des permutations des rangs des produits associés à chaque sujet de façon à optimiser le critère suivant :

$$\lambda \sum_{i,j} r_{ij}^2 + (1-\lambda) \sum_{i,j} s_{ij}^2$$

Où les r_{ij} sont les éléments de la matrice des fréquences des positions et les s_{ij}, les éléments de la matrice de carry-over. Dès que l'optimum ou que le nombre maximum d'itérations sont atteints, l'algorithme s'arrête.

XLSTAT utilise deux indices afin de vérifier la qualité de ces deux matrices :

- le MDR (mean deviation of R) : $MDR = \sum_{i,j}(r_{ij} - \overline{r})$, c'est-à-dire le déviation par rapport à la moyenne des éléments de la matrice des fréquences des positions.

- le MDS (mean deviation of S): $MDS = \sum_{i,j}(s_{ij} - \overline{s})$, c'est-à-dire le déviation par rapport à la moyenne des éléments de la matrice du carry-over.

Pour les plans en blocs incomplets équilibrés, on peut calculer la valeur optimale de ces indices ce qui permet de voir si le plan obtenu est optimal.

Plans résolvables et présentation améliorée

Un plan résolvable est un plan qui peut être subdivisé en g groupes de sujets tels que qu'à l'intérieur de chaque groupe, on ait une occurrence de chaque produit. Certains plans en blocs incomplets équilibrés possèdent cette propriété. Présenter un plan avec une telle subdivision en groupes présente l'avantage que si certains sujets ne se présentent pas, il n'est pas nécessaire de reconstruire un plan d'expériences, mais simplement de faire en sorte que les expériences annulées soient les dernières du plan. Cette approche est aussi particulièrement intéressante lorsque l'on veut mettre en place plusieurs sessions d'évaluation (voir ci-dessous). Une condition pour qu'un plan en blocs incomplets équilibrés soit résolvable est que $\frac{v}{k}$ doit être un entier.

Même lorsqu'ils ne sont pas équilibrés ou résolvables, XLSTAT cherche à présenter les plans en blocs incomplets de telle manière que les produits soient présents au plus deux fois et si possible une seule fois dans un groupe de taille $< \frac{v}{k} >$ (où $< i >$ vaut i si i est un entier, et

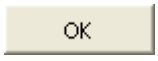
l'arrondi à la valeur entière supérieure sinon). Ainsi, si des sujets étaient finalement absents, cela ne pénaliserait pas trop la qualité du plan tel qu'il a été initialement conçu.

Sessions

Il peut arriver que dans une expérimentation on ait trop de produits à tester par sujet et qu'il faille plusieurs sessions afin de ne pas saturer les sujets. Lorsque plusieurs sessions sont demandées, XLSTAT utilise le même plan initial pour chaque session, puis effectue une permutation des sujets et des positions à l'intérieur de groupes de sujets, à chaque session. Pour les plans résolvables, les groupes de sujets utilise des groupes de taille v/k. Ainsi, dans la mesure du possible, un même sujet n'évaluera pas deux fois le même produit d'une session à l'autre.

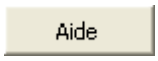
Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.



: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.

Onglet **Général** :

Produits : entrez le nombre de produits qui seront évalués par l'ensemble des sujets.

Produits / Sujet : entrez le nombre de produits que devra évaluer chaque sujet. Si vous activez l'option session, entrez le nombre de produits que devra évaluer chaque sujet au cours de chaque session.

Sujets : entrez le nombre de sujets évaluant les produits.

Sessions : activez cette option si plusieurs sessions sont envisagées. Si tel est le cas, entrez le nombre de session prévues.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des sujets : activez cette option si vous voulez utiliser des libellés pour les sujets pour l'affichage des résultats.

Onglet **Options** :

Méthode : choisissez la méthode de recherche du plan d'expérience.

- **Rapide** : activez cette option pour utiliser une méthode rapide limitant au maximum le temps de recherche.
- **Recherche** : activez cette option pour définir le temps alloué à la recherche de l'optimum. Le temps maximum doit être défini en secondes.

Critère : choisissez le critère à maximiser lors de la recherche du plan d'expérience.

- **A-efficacité** : activez cette option pour rechercher un plan maximisant le critère d'A-efficacité.
- **D-efficacité** : activez cette option pour rechercher un plan maximisant le critère de D-efficacité.

Effet de report vs effet d'ordre : définissez ici vos préférences sur ce qui doit être la priorité de XLSTAT dans la seconde phase de génération du plan d'expériences : l'homogénéité de la fréquence des ordres d'évaluation des produits (**effet d'ordre ou de position**), ou l'homogénéité des nombre de fois où deux produits se succèdent pour l'ensemble des sujets (**effet de report ou carry-over**).

- **Lambda** : faites varier ce critère entre 0 (priorité carry-over) et 1 (priorité fréquences des positions).
- **Itérations** : entrez le nombre maximal d'itérations à autoriser pour l'algorithme de recherche de la meilleure solution.

Codes produits : choisissez l'option pour la génération des codes produits.

- **Identifiant produit** : activez cette option pour utiliser simple identifiant produit (1,2, …).
- **Code aléatoire** : activez cette option pour utiliser un code aléatoire généré par le logiciel.

- **Défini par l'utilisateur** : activez cette option pour sélectionner sur une feuille de calcul une colonne contenant les noms des produits impliqués dans le plan d'expérience. Le nombre de produits sélectionnés doit correspondre au nombre de produit défini plus haut pour les plan d'expériences.

Onglet **Sorties** :

Tableau Sujets x Produits : activez cette option pour afficher le tableau binaire indiquant si un sujet a évalué (valeur 1) ou non (valeur 0) un produit.

Tableau des cooccurrences : activez cette option pour afficher le tableau des cooccurrences indiquant combien de fois deux produits ont été évalués par un même sujet.

Tableau Sujets x Rangs : activez cette option pour afficher le tableau indiquant, pour chaque sujet, quel produit est évalué à chaque étape de l'expérience.

Tableau des effets d'ordre : activez cette option pour afficher le tableau indiquant combien de fois chaque produit a été vu à chaque étape de l'expérience.

Tableau des carry-over : activez cette option pour afficher le tableau indiquant combien de fois chaque produit a été évalué juste après un autre.

Tableau du plan d'expériences : activez cette option pour afficher le tableau qui pourra être analysé avec une ANOVA, une fois les évaluations recueillies.

Résultats

Une fois les calculs terminés, XLSTAT indique le temps passé à la recherche du plan optimal. Les deux critères A-efficacité et D-efficacité sont affichés. Si le plan optimal a été trouvé (cas d'un plan en blocs incomplets équilibrés) XLSTAT l'indique. De même, si le plan est résolvable, cela est indiqué et la taille des groupes est précisée.

Si des sessions ont été demandées, une première série de résultats est affichée avec les tableaux prenant en compte l'ensemble des sessions. Les résultats correspondant à chaque session sont ensuite affichés.

Le premier tableau présenté est le **tableau Sujets x Produits** indiquant si un sujet a évalué (valeur 1) ou non (valeur 0) un produit.

Le **tableau des cooccurrences** indique combien de fois deux produits ont été évalués par un même sujet.

Le **tableau MDS/MDR** donne des indices qui permettent de juger de la qualité des rangs obtenus. Dans ce tableau apparaît les valeurs optimales lorsqu'elles peuvent être calculées et les valeurs obtenues sur le plan.

Le **tableau Sujets x Rangs** indique, pour chaque sujet, quel produit est évalué à chaque étape de l'expérience.

Le **tableau des effets d'ordre** indique combien de fois chaque produit a été vu à chaque étape de l'expérience.

Le **tableau des carry-over** indique combien de fois chaque produit a été évalué juste après un autre.

Le **tableau du plan d'expériences** pourra être analysé avec une ANOVA, une fois les évaluations recueillies.

Exemple

Un exemple de génération de plan d'expérience est disponible sur le Centre d'aide XLSTAT :

https://www.xlstat.com/demo-doesensof.htm

Bibliographie

John J.A.and Whitaker D. (1993). Construction of cyclic designs using integer programming. *Journal of Statistical Planning and Inference*, **36** , 357-366.

John J.A.and Williams E.R. (1995). Cyclic Designs and Computer-Generated Designs. New York, Chapman & Hall.

Périnel E. and Pagès J. (2004). Optimal nested cross-over designs in sensory analysis. *Food Quality and Preference*, **15** (5), 439-446.

Wakeling I.N, Hasted A. and Buck D. (2001). Cyclic presentation order designs for consumer research. *Food Quality and Preference*, **12**, 39-46

Williams E.J. (1949). Experimental designs balanced for the estimation of residual effects of treatments. *Aust. J. of Sci. Res.*, **2**, 149-164.

Plans d'expériences pour les tests de discrimination sensorielle

Cet outil vous permet de générer des plans d'expériences pour mettre en place des tests de discrimination sensorielle. Il permet de générer des plans pour les tests du triangle, duo-trio, deux parmi cinq, 2-AFC, 3-AFC et des tétrades.

Dans cette section :

Description

La planification expérimentale est une étape fondamentale pour quiconque veut s'assurer que les données collectées seront exploitables dans les meilleures conditions statistiques possibles. Rien ne sert de faire évaluer des produits par un panel de sujets si l'on ne peut ensuite comparer les produits dans des conditions statistiques satisfaisantes.

Cet outil a pour but de permettre aux spécialistes de l'analyse sensorielle de disposer d'un outil simple et puissant pour mettre en place un test de discrimination sensorielle pour évaluer un ensemble de produits.

Dans la mise en place de nouveaux produits en analyse sensorielle, les tests de discrimination prennent une place très importante. XLSTAT permet de générer les combinaisons de produits à présenter aux sujets afin d'effectuer ce type de tests. Il permet ensuite d'analyser les résultats.

Les tests de discrimination se basent sur la différenciation de deux produits présentés de manières différentes en fonction du test choisi.

Les informations requises pour générer ce type de plans d'expérience sont le type de test, le nombre de sujets et si possible le nom des produits.

Les tests disponibles dans XLSTAT sont :

- le test du triangle : 3 produits sont présentés à chaque sujet dans des ordres différents. Parmi ces 3 produits, 2 sont similaires et le troisième est différent. Les sujets doivent identifier le produit différent des deux autres.
- le test duo-trio : les sujets testent un produit de référence, puis ils testent deux produits dont un est le produit de référence. Les sujets doivent identifier le produit de référence.
- le test deux parmi cinq : 5 produits sont présentés à chaque sujet. Les produits sont de 2 types, répartis en deux groupes, l'un de 3 produits et l'autre de 2 produits. Les sujets

doivent identifier les 2 produits similaires appartenant au groupe de 2.

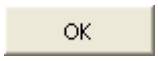
- le test 2-AFC : 2 produits sont présentés à chaque sujet. Chaque sujet doit identifier le produit ayant l'intensité la plus forte pour une caractéristique spécifique.
- le test 3-AFC : 3 produits sont présentés à chaque sujet, deux similaires et un différent des deux autres. Chaque sujet doit identifier le produit ayant l'intensité la plus forte pour une caractéristique spécifique.
- le test des tétrades : 4 produits identiques deux par deux sont présentés aux sujets. Chaque sujet doit identifier les deux groupes de produits.

Pour chaque test, nous pouvons générer un plan d'expérience obtenu par randomisation des combinaisons possibles.

L'utilisateur peut spécifier le nombre de sessions et des libellés pour les sujets et les produits.

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.


: cliquez sur ce bouton pour changer la façon dont vous chargez les données dans XLSTAT. Si le bouton possède une icône de souris, XLSTAT vous permet de faire une sélection à la souris des données. Si le bouton possède une icône de liste, XLSTAT vous permet de sélectionner les données à partir d'une liste. Si le bouton possède une icône orange, des boutons complémentaires avec un point d'interrogation  sont affichés pour vous permettre d'importer les données à partir de fichiers.

Onglet **Général**:

Type de test : sélectionnez le test que vous désirez mettre en place.

Sujets : entrez le nombre de sujets évaluant les produits.

Sessions : activez cette option si plusieurs sessions sont envisagées. Si tel est le cas, entrez le nombre de sessions prévues.

Libellés des sujets : activez cette option si vous voulez utiliser des libellés pour les sujets pour l'affichage des résultats.

Libellé inclus : activez cette option si la première ligne des données sélectionnées contient un libellé.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Codes produits : choisissez l'option pour la génération des codes produits.

- **Identifiant produit** : activez cette option pour utiliser un identifiant produit simple (P1, P2, …).
- **Code aléatoire** : activez cette option pour utiliser un code aléatoire généré par le logiciel. Deux choix sont disponibles : le premier, *Alphabétiques*, permet de générer un code aléatoire avec trois lettres et le second, *Numériques* permet de générer un code aléatoire avec trois chiffres.
- **Défini par l'utilisateur** : activez cette option pour sélectionner sur une feuille de calcul deux colonnes (une pour chaque produit) contenant les noms des produits impliqués dans le plan d'expérience. Le nombre de lignes sélectionnées doit correspondre au nombre d'échantillons différents nécessaires au plan d'expérience. Ainsi, il faut 2 lignes pour les tests du triangle, duo-trio, 3-AFC et des tétrades. Le test 2-AFC n'en nécessite qu'une et le test deux parmi cinq en nécessite 3. Enfin, il faut noter que dans le cas des tests duo-trio et 3-AFC, la cellule (2,2) des données sélectionnées doit être vide puisque le produit 2 ne nécessite qu'un échantillon.

Résultats

Une fois les calculs terminés, XLSTAT affiche la question à poser aux sujets liée au test sélectionné.

Le tableau des produits à tester pour chaque sujet est ensuite affiché, la dernière colonne pourra être remplie suite au test afin de lancer l'analyse. On peut noter que pour les tests du triangle, duo-trio, 2-AFC et 3-AFC, l'utilisateur peut rentrer les réponses de chaque sujet dans le tableau et la réponse correcte/incorrecte sera automatiquement remplie. Pour le 2-AFC, il faudra rentrer la réponse correcte (qui est la même pour toute la colonne) pour utiliser cette fonctionnalité.

Exemple

Un exemple de génération de plan d'expérience et de traitement d'un test de discrimination sensoriel est disponible sur le Centre d'aide XLSTAT :

http://www.xlstat.com/demo-testsensof.htm

Bibliographie

Bi J. (2008). Sensory discrimination tests and measurements: Statistical principles, procedures and tables. John Wiley & Sons.

Næs T., Brockhoff P. B., and Tomiæ O. (2010). Statistics for Sensory and Consumer Science. John Wiley & Sons, Ltd.

Tests de discrimination sensorielle

Cet outil vous permet d'appliquer les tests de discrimination en analyse sensorielle suite à la mise en place de tests de différents types comme le test du triangle, le test des tétrades, le test duo-trio, le test 2-AFC, le test 3-AFC et le test deux parmi cinq. Il vous permet d'obtenir des résultats sur la significativité de ces tests sensoriels ainsi que sur leur puissance.

Dans cette section :

Description

Dans la mise en place de nouveaux produits en analyse sensorielle, les tests de discrimination prennent une place importante. XLSTAT permet d'analyser le résultat de ces tests en utilisant différentes méthodes.

Trois modèles d'estimation sont disponibles, en premier l'approche dite « *guessing approach* », en second, l'approche issue des concepts de Thurstone que l'on appellera Thurstonienne, et enfin une troisième approche pour les tests qui se déroulent sur plusieurs sessions, qui est basée sur le modèle Beta-Binomial.

Les tests disponibles dans XLSTAT sont :

- le test du triangle : 3 produits sont présentés à chaque sujet dans des ordres différents. Parmi ces 3 produits, 2 sont similaires et le troisième est différent. Les sujets doivent identifier le produit différent des deux autres.
- le test duo-trio : les sujets testent un produit de référence, puis ils testent deux produits dont un est le produit de référence. Les sujets doivent identifier le produit de référence.
- le test deux parmi cinq : 5 produits sont présentés à chaque sujet. Les produits sont de 2 types, répartis en deux groupes, l'un de 3 produits et l'autre de 2 produits. Les sujets doivent identifier les 2 produits similaires appartenant au groupe de 2.
- le test 2-AFC : 2 produits sont présentés à chaque sujet. Chaque sujet doit identifier le produit ayant l'intensité la plus forte pour une caractéristique spécifique.
- le test 3-AFC : 3 produits sont présentés à chaque sujet, deux similaires et un différent des deux autres. Chaque sujet doit identifier le produit ayant l'intensité la plus forte pour une caractéristique spécifique.
- le test des tétrades : 4 produits identiques deux par deux sont présentés aux sujets. Chaque sujet doit identifier les 2 groupes de produits.

Ces tests ont chacun des avantages et des inconvénients qui ont été analysés par Bi (2008).

Quelques concepts doivent être introduits : la probabilité d'obtenir une réponse correcte pC, la probabilité de discrimination pD, la probabilité de deviner pG et le d', appelé d-prime ou delta de Thurstone.

Modèles

Deux modèles sont couramment utilisés pour les tests de discrimination sensorielle.

Le modèle basé sur le hasard (guess) suppose que les consommateurs sont soit des individus discriminants, soit des individus non-discriminants. Les individus discriminants arrivent toujours à différencier les produits. Les non-discriminants se basent sur le hasard pour différencier les produits. Par exemple, ils ont une probabilité de 1/3 de trouver la bonne réponse avec le test du triangle. La proportion de discriminateurs est la proportion d'individus qui décèlent la différence entre les produits.

Ce concept peut s'exprimer de cette façon : $pD = \frac{(pC - pG)}{(1 - pG)}$ où pC est la probabilité d'une réponse correcte et pG la probabilité d'obtenir la bonne réponse au hasard.

Dans le modèle de Thurstone, on n'utilise pas une probabilité pD mais une mesure d' (d-prime). Il s'agit d'une distance sensorielle entre 2 produits. Une unité représente un écart-type.

Dans ce cas, l'hypothèse du modèle est que les représentations sensorielles des produits suivent une loi normale et que les individus ne peuvent pas être catégorisés en discriminants / non-discriminants. On suppose que le consommateur a toujours raison dans ce qu'il ressent. Une réponse incorrecte ne traduit pas une erreur du sujet mais une forte ressemblance entre les produits. Si d' est proche de 0, les produits ne peuvent pas être différenciés.

Pour chaque test, nous avons donc une probabilité d'obtenir la bonne réponse au hasard mais aussi une fonction psychométrique spécifique à chaque test qui permettra de relier la probabilité d'une réponse correcte au d'. Tous ces paramètres sont spécifiques à chaque test. Nous avons $pC = f_{\text{test}}(d')$.

Probabilité en cas de réponse au hasard

Pour chaque test, la probabilité d'une réponse correct en cas de réponse au hasard, est égale à :

Test du triangle : $pG = 1/3$

Test duo-trio : $pG = 1/2$

Test deux parmi cinq : $pG = 1/10$

2-AFC : $pG = 1/2$

3-AFC : $pG = 1/3$

Test des tétrades : $pG = 1/3$

Fonctions psychométriques

Pour chaque test la fonction psychométrique qui relie la probabilité d'une réponse correcte au d' est la suivante :

Test du triangle : $pC = f_{triangle}(d') = 2\int_0^\infty \{\Phi[-x\sqrt{3} + d'\sqrt{2/3}] + \Phi[-x\sqrt{3} - d'\sqrt{2/3}]\}\phi(x)dx$

Test duo-trio : $pC = f_{duo-trio}(d') = -\Phi(d\prime/\sqrt{2}) - \Phi(d\prime/\sqrt{6}) + \Phi(d\prime/\sqrt{2})\Phi(d\prime/\sqrt{6})$

2-AFC : $pC = f_{2-AFC}(d') = \Phi(d'/\sqrt{2})$

3-AFC : $pC = f_{3-AFC}(d') = \int_{-\infty}^{\infty} \phi(x - d')\Phi[x]^2 dx$

Test des tétrades : $pC = f_{\text{tetrad}}(d') = \int_{-\infty}^{\infty} \phi(x)\Phi[x]\{1 - \Phi[x - d']\}^2 dx$

Les intégrales sont estimées en utilisant la méthode de Gauss-Kronrod.

Calcul des p-valeurs et de la puissance

Les p-valeurs et la puissance sont obtenues en utilisant soit la distribution binomiale soit la distribution normale basée sur la pC estimée.

Ecart-type et intervalle de confiance pour les paramètres du modèle de Thurstone

Lorsque le modèle de Thurstone est utilisé, on peut obtenir des écarts-types et des intervalles de confiance pour les paramètres d'intérêt.

Pour la probabilité d'une réponse correcte, nous avons :

$$SE(pC) = \sqrt{pC(1 - pC)/N}$$

avec N le nombre de sujets.

Pour la probabilité de discrimination, nous avons :

$$SE(pD) = \frac{SE(pC)}{1 - pG}$$

Pour le d', nous avons :

$$SE(d') = \frac{SE(pC)}{f'_{test}(d')}$$

Où f' est la dérivée de la fonction psychométrique par rapport à d' (Brockhoff and Christensen, 2010).

Modèle Beta-binomial

Le modèle binomial, utilisé pour le modèle basé sur le hasard et le modèle thurstonien, repose sur les hypothèses que les choix de chaque sujet sont indépendants et que la probabilité d'une

réponse correcte est la même pour tous. Cependant, lorsque des sessions sont introduit dans les données, ces hypothèses ne sont plus respectées. On parle de sessions lorsqu'un test de discrimination sensorielle effectué sur k sujets, est répété sur l'ensemble ou une partie des sujets jsuqu'à n fois. Dans ce cas, l'indépendance entre les réponses n'est plus respectée car elles peuvent provenir d'un même sujet, ainsi la probabilité de choisir la réponse correcte varie suivant les sujets en fonction de leur capacité de discrimination. On parle alors de surdispersion car les données comportent plusieurs sources de variation.

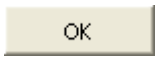
Le modèle Beta-Binomial permet de prendre en compte ce phénomène dans les résultats des tests par le biais de l'estimation des paramètres μ et $gamma$. L'estimation de μ représente celle de la probabilité d'une réponse correcte pC. Quant au paramètre $gamma$, il permet de mesurer la variation entre les sujets. Si $gamma$ est proche de 0, cela signifie qu'il n'y a pas de surdispersion cela revient à utiliser le modèle binomial. Au contraire, si $gamma$ est proche de 1, le modèle Beta-Binomial doit être utilisé pour respecter les hypothèses sous-jacentes, et ainsi éviter une sous estimation des écarts-types et une interprétation erronée.

Les paramètres du modèle Beta-Binomial sont estimés par maximum de vraisemblance.

A travers le modèle Beta-Binomial, nous allons tester à la fois si les probabilités d'une réponse correcte de chaque sujet sont égales à la probabilité d'obtenir une bonne réponse au hasard pG, ainsi que l'existence d'une variation entre les sujets. Dès qu'une des deux hypothèses de test est rejetée ($\mu = pG$ ou $gamma = 0$), les produits sont considérés comme différents par les sujets.

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.


: cliquez sur ce bouton pour changer la façon dont vous chargez les données dans XLSTAT. Si le bouton possède une icône de souris, XLSTAT vous permet de faire une sélection

à la souris des données. Si le bouton possède une icône de liste, XLSTAT vous permet de sélectionner les données à partir d'une liste. Si le bouton possède une icône orange, des boutons complémentaires avec un point d'interrogation  sont affichés pour vous permettre d'importer les données à partir de fichiers.

Onglet **Général** :

Type de test : choisissez le test discriminatif à utiliser.

Méthode : sélectionnez la méthode à utiliser en fonction du modèle (guessing model ou modèle de Thurstone). Pour utiliser le modèle Beta-Binomial, il faut choisir l'option *Données avec sessions* dans la liste déroulante des *Données en entrée*.

Données en entrée : choisissez le type de données en entrée. Quatre types sont disponibles et les options affichées varient en fonction de ce choix.

- **Cas de la sélection de données**

Résultats du test : sélectionnez une colonne dans laquelle on différencie un sujet ayant identifié les différences et un sujet n'ayant pas identifié les différences.

Code pour une réponse correcte : entrez le code utilisé pour coder une réponse correcte du sujet.

- **Cas de la taille d'échantillon :**

Nombre de sujets : entrez le nombre total de sujets pour ce test.

Nombre de réponses correctes : entrez le nombre de réponses correctes obtenues suite au test.

- **Cas de la proportion :**

Nombre de sujets : entrez le nombre total de sujets pour ce test.

Proportion de réponses correctes : entrez la proportion de réponses correctes obtenues suite au test.

- **Cas de la sélection des données avec sessions** :

Tableau avec sessions : sélectionnez un tableau comprenant 2 colonnes. La première colonne contient le nombre de réponses correctes obtenues pour chaque sujet. La deuxième colonne contient le nombre de sessions auxquel a participé chaque sujet.

Les options suivantes apparaissent dans le cas où le modèle de Thurstone a été sélectionné. Elles permettent de préciser les hypothèses du test qui doivent être vérifiées.

Hypothèse nulle :

D-prime : activez cette option si vous voulez effectuer un test sur une valeur de d'. Vous pouvez ensuite entrer la valeur souhaitée. L'hypothèse nulle du test sera alors " d' *est égal à* x ", avec x la valeur choisie.

pD : activez cette option si vous voulez effectuer un test sur une valeur de la probabilité de discrimination. Vous pouvez ensuite entrer la valeur souhaitée. L'hypothèse nulle du test sera alors " *La probabilité de discrimination est égale à* x ", avec x la valeur choisie.

Remarque : par défaut, l'hypothèse de test qui est vérifiée est si d' est égal à 0.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des colonnes : activez cette option si la première ligne des données sélectionnées contient un libellé.

Niveau de signification (%) : entrez le niveau de signification à utiliser pour le test (valeur par défaut : 5%).

Statistique : sélectionnez la méthode de calcul pour les intervalles de confiance.

Puissance : sélectionnez la distribution utilisée pour le calcul de la puissance.

Remarque : dans le cas du modèle Beta-Binomial, les intervalles de confiance sont calculés suivant la méthode de Wald.

Résultats

Récapitulatif des options sélectionnées : dans ce tableau sont affichés les paramètres sélectionnés dans la boîte de dialogue.

Test de discrimination sensorielle : ce tableau regroupe les résultats du test réalisé, en commençant par la valeur estimée du paramètre d'intérêt, à savoir d-prime ou pD pour le modèle de Thurstone, et μ (mu) et $gamma$ pour le modèle Beta-Binomial. L'interprétation du test, ainsi qu'une p-value et une puissance sont indiquées.

Paramètres estimés : dans ce tableau sont affichés les probabilités, le d-prime, les paramètres μ et $gamma$ du modèle Beta-Binomial le cas échéant, ainsi que leurs écarts-types et les intervalles de confiance associés. Ce tableau n'est pas affiché dans le cas du guessing model.

Exemple

Un premier exemple de test de discrimination en analyse sensorielle est disponible sur le Centre d'aide XLSTAT à l'adresse :

http://www.xlstat.com/demo-testsensof.htm

et un autre exemple avec des données comportant des sessions est disponible à l'adresse :

...

Bibliographie

Bi J. (2008). Sensory discrimination tests and measurements: Statistical principles, procedures and tables. John Wiley & Sons.

Bi J. and O'Mahony M. (2013). Variance of *d'* for the Tetrad Test and Comparisons with Other Forced-Choice Methods. *Journal of Sensory Studies*, **28**, 91-101.

Brockhoff, P.-B., Christensen, R. H. B. (2010). Thurstonian models for sensory discrimination tests as generalized linear models, *Food Quality and Preference*, **21**, 330-338.

Kunert, J., and Meyners, M. (1999). On the triangle test with replications. *Food Quality and preference*, **10(6)**, 477-482.

Liggett, R. E., and Delwiche, J. F. (2005). The beta-binomial model: Variability in overdispersion across methods and over time. *Journal of Sensory Studies*, **20(1)**, 48-61.

Næs T., Brockhoff P. B., and Tomiæ O. (2010). Statistics for Sensory and Consumer Science. John Wiley & Sons, Ltd.

Puissance- Tests de discrimination sensorielle

Cet outil vous permet de contrôler votre puissance ou votre nombre de sujets dans le cadre des tests de discrimination sensorielle. Les tests de discrimination inclus sont les tests du triangle, duo-trio, deux parmi cinq, 2-AFC, 3-AFC et des tétrades.

Cette fonction permet de :

- déterminer la puissance en fonction du nombre de sujets
- déterminer le nombre de sujets en fonction de la puissance

Dans cette section :

Description

La planification expérimentale est une étape fondamentale pour quiconque veut s'assurer que les données collectées seront exploitables dans les meilleures conditions statistiques possibles. Rien ne sert de faire évaluer des produits par un panel de sujets si l'on ne peut ensuite pas comparer les produits dans des conditions statistiques satisfaisantes.

Cet outil a pour but de permettre aux spécialistes de l'analyse sensorielle de disposer d'un outil simple et puissant précédant la mise en place un test de discrimination sensorielle afin d'évaluer un ensemble de produits.

Lorsque nous testons une hypothèse à l'aide d'un test statistique,nous avons plusieurs éléments à choisir :

- L'hypothèse nulle H0 et l'hypothèse alternative Ha
- Le test statistique à utiliser
- L'erreur de première espèce (erreur de type I) que l'on appelle aussi $alpha$. Elle se produit lorsqu'on rejette l'hypothèse nulle alors que celle-ci est vraie. Elle est fixée a priori pour chaque test et vaut 5%.

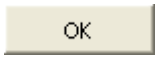
L'erreur de seconde espèce ou beta est moins étudiée mais elle revêt une grande importance. En effet, elle représente la probabilité que l'on ne rejette pas l'hypothèse nulle alors que celle-ci est fausse. On ne peut pas la fixer a priori mais, on peut essayer de la minimiser, en jouant sur les autres paramètres du modèle. La puissance d'un test est calculée comme 1-beta et représente la probabilité que l'on rejette l'hypothèse nulle alors que celle-ci est bien fausse.

On voudra donc maximiser la puissance du test. XLSTAT permet de calculer cette puissance lorsque les autres paramètres du test sont connus. D'autre part, il permet pour une puissance donnée d'évaluer la taille de l'échantillon (le nombre de sujets) nécessaire à l'obtention de cette puissance.

Pour plus de détails sur les tests de discrimination sensorielle, se référer aux Plans d'expériences pour les tests de discrimination sensorielle ainsi qu'aux Tests de discrimination sensorielle.

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.

Onglet **Général** :

Objectif : choisissez entre le calcul de la puissance et la recherche du nombre d'observations (en fonction de ce choix les champs suivants seront différents).

Type de test : sélectionnez le test à utiliser.

Alpha : entrez l'erreur de première espèce.

Puissance (dans le cas où l'on recherche la taille de l'échantillon) : entrez la puissance requise.

Taille d'échantillon (dans le cas où l'on cherche à calculer la puissance) : entrez la taille de l'échantillon (le nombre de sujets).

Hypothèse nulle :

D-prime : activez cette option si vous voulez effectuer un test sur une valeur de d'. Vous pouvez ensuite entrer la valeur souhaitée. L'hypothèse nulle du test sera alors " d' *est égal à* x ", avec x la valeur choisie.

pD : activez cette option si vous voulez effectuer un test sur une valeur de la probabilité de discrimination. Vous pouvez ensuite entrer la valeur souhaitée. L'hypothèse nulle du test sera alors " *La probabilité de discrimination est égale à* x ", avec x la valeur choisie.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Puissance : sélectionnez la distribution utilisée pour le calcul de la puissance.

Résultats

Dans un premier temps, un tableau de résultats est affiché. Il est composé de 2 colonnes: la proportion de réponses correctes, suivie de la taille de l'échantillon (ou la puissance en fonction des paramètres sélectionnés dans la boîte de dialogue). Il permet de construire le graphique de simulation qui est affiché en-dessous.

Exemple

Un exemple de calcul de la taille d'échantillon nécessaire est disponible sur le Centre d'aide XLSTAT à l'adresse

https://www.xlstat.com/demo/trp_fr

Bibliographie

Brockhoff, P.B. and Christensen, R.H.B (2010). Thurstonian models for sensory discrimination tests as generalized linear models. *Food Quality and Preference*, **21**, 330-338.

Ennis, J.M. and V. Jesionka (2011). The power of sensory discrimination methods revisited. *Journal of Sensory Studies*, **26**, 371-382.

Créer un tableau Produits\Sujets

Utilisez cet outil pour transformer vos données sensorielles verticales en un tableau Produits/Sujets (horizontal).

Dans cette section :

Description

Les données sensorielles peuvent être présentées de deux manières :

1. **Forme verticale** : il s'agit d'un tableau comprenant **autant de lignes que d'échantillons**. Une colonne est dédiée à l'identifiant du produit, une autre à l'identifiant du sujet, et les colonnes restantes correspondent aux descripteurs tels que la douceur, le sucré, etc. La taille du tableau est déterminée par la formule $n \times p + 2$, où n représente le nombre d'échantillons et p le nombre de variables descriptives.

2. **Forme horizontale** : il s'agit d'un tableau comprenant **autant de lignes que de produits**. Le nombre de colonnes est égal au produit du nombre de sujets par le nombre de descripteurs. La taille du tableau est donc $n \times (p \times m)$, avec n désignant le nombre de produits, p le nombre de variables descriptives et m le nombre de sujets. Dans la forme horizontale, les données du premier sujet sont affichées dans les p premières colonnes, les données du deuxième sujet sont affichées dans les p colonnes suivantes, et ainsi de suite.

Un **tableau Produits/Sujets**, qui représente les données sensorielles sous une forme horizontale, est nécessaire pour certaines analyses sensorielles de XLSTAT :

- Analyse Factorielle Multiple (MFA) ;
- Analyse Procrustéenne Généralisée (GPA) ;
- STATIS ;
- CLUSTATIS ;
- Analyse de données de projective mapping.

Cependant, il se peut que les données sensorielles ne soient disponibles qu'en forme verticale. Cette fonctionnalité permet de convertir les données sensorielles verticales en un **tableau Produits/Sujets**.

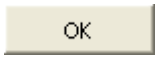
Gestion des sessions

La fonctionnalité **Créer un tableau Produits/Sujets** gère automatiquement les sessions (ou répétitions), dans le cas où un même sujet goûte ou teste plusieurs fois un même produit.

Lorsqu'une combinaison produit/sujet est identifiée plusieurs fois, la moyenne des données pour chaque descripteur est calculée.

Boîte de dialogue

Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT vous permet de sélectionner les données par colonnes ou par plage. Si la flèche est vers la droite, XLSTAT vous permet de sélectionner les données par lignes ou par plage.


: cliquez sur ce bouton pour changer la façon dont vous chargez les données dans XLSTAT. Si le bouton possède une icône de souris, XLSTAT vous permet de faire une sélection à la souris des données. Si le bouton possède une icône de liste, XLSTAT vous permet de sélectionner les données à partir d'une liste. Si le bouton possède une icône orange, des boutons complémentaires avec un point d'interrogation  sont affichés pour vous permettre d'importer les données à partir de fichiers.

Données : sélectionnez les colonnes correspondant aux descripteurs. Si des en-têtes de colonnes ont été sélectionnés, veuillez vérifier que l'option « Libellés des variables » est activée.

Produits : sélectionnez la colonne contenant les identifiants des produits. Si des en-têtes de colonnes ont été sélectionnés, veuillez vérifier que l'option « Libellés des variables » est activée.

Sujets : sélectionnez la colonne contenant les identifiants des sujets. Si des en-têtes de colonnes ont été sélectionnés, veuillez vérifier que l'option « Libellés des variables » est activée.

Plage : activez cette option pour que les résultats soient affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des variables : activez cette option si la première ligne des données sélectionnées contient un libellé.

Afficher l'en-tête du rapport : désactivez cette option si vous souhaitez que le tableau des résultats commence dès la première ligne de la feuille Excel (cas d'une sortie dans une feuille ou un classeur), et non après l'en-tête du rapport.

Afficher les libellés des sujets : activez cette option pour afficher les libellés des sujets en haut du tableau Produits/Sujets.

Résultats

En activant l'option **Afficher l'en-tête du rapport** plusieurs informations descriptives sont affichées :

- les différentes sélections données, produits et sujets ;
- le nombre de sujets identifiés ;
- le nombre de produits identifiés ;
- le nombre de sessions.

Le **tableau Produits/Sujets** est affiché et présente les éléments suivants :

- la première ligne affiche les libellés des **sujets** (si l'option a été activée) ;
- la deuxième ligne affiche les libellés des descripteurs, qui correspondent aux mêmes libellés entrés dans le champ **Données** ;
- la première colonne contient les libellés des **produits** ;
- enfin, le corps du tableau reprend les différentes données triées par sujets.

Exemple

Un tutoriel pour créer un tableau Produits/Sujets est disponible sur le Centre d'aide XLSTAT :

http://www.xlstat.com/demo-sdp.htm

JAR analyse multivariée et classification

Utilisez cet outil pour faire une analyse multivariée (CATATIS) ou une classification (CLUSCATA) sur des données de type JAR (*Just About Right*).

Dans cette section :

Description

Les données JAR sont bien souvent traitées uniquement avec l'analyse de pénalité mais regorgent d'autres informations. En effet, tout comme les autres épreuves, elles permettent de décrire les produits, de voir les ressemblances et différences, etc. Ainsi, faire une analyse multivariée, permettant à la fois de créer une carte des produits avec leurs descriptions, est très instructif.

Données JAR

Les données JAR sont des données mesurées sur une échelle JAR (*Just About Right*) sur 5, 7 ou 9 niveaux. Dans le cas de 5 niveaux, ces données correspondent à des notes de 1 à 5 pour une ou plusieurs caractéristiques des produits étudiés où 1 correspond à « Pas du tout assez », 2 à « Pas assez », 3 à « JAR » un idéal pour le consommateur, 4 à « Trop » et 5 à « Beaucoup trop ». Par exemple, pour un chocolat, on pourra noter son amertume, et pour le confort d'une voiture, le volume sonore du moteur.

Analyse CATATIS et CLUSCATA sur des données JAR

En analysant des données JAR avec CATATIS et CLUSCATA, il est alors possible (Llobell, 2022) :

- d'étudier les liens entre les produits et les attributs ;
- d'analyser l'homogénéité des réponses, ce qui est très informatif sur la qualité de vos données ;
- de construire automatiquement des groupes de sujets ayant différent point de vue grâce à une amélioration de la méthode CLUSCATA.

Pour utiliser des analyses sensorielles telles que CATATIS et CLUSCATA sur des données JAR, il faut tout d'abord prétraiter ses données. XLSTAT propose alors de lancer une analyse CATATIS ou CLUSCATA à partir de données JAR sans avoir à prétraiter ses données au préalable.

Cosinus de Salton

L'accord entre deux sujets est calculé à l'aide du cosinus de Salton qui est équivalent à l'indice d'Ochiai (Salton & McGill, 1983) dans le cas de données binaires (Llobell, 2022) :

$$s(X,Y) = \frac{< X,Y >}{||X||||Y||} = \frac{trace(XY^T)}{\sqrt{(trace(XX^T)trace(YY^T))}},$$

où 0 correspond à un désaccord total et 1 à un accord parfait.

Prétraitement sur les données JAR

Pour ce faire, cet outil commence par transformer les données JAR en un tableau horizontal :

- Les données originales comprennent un sujet, trois produits et un attribut :

	Attribut 1
Produit 1	Pas assez
Produit 2	JAR
Produit 3	Trop

- Une transformation du tableau précédant en tableau disjonctif est appliquée :

	Pas assez	**JAR**	**Trop**
Produit 1	**1**	0	0
Produit 2	0	**1**	0
Produit 3	0	0	**1**

- Enfin un codage flou (*Fuzzy coding*) est appliqué aux valeurs pour chaque produit, afin de tenir compte de l'ordinalité des données :

	Pas assez	**JAR**	**Trop**
Produit 1	**1-**β	β	0
Produit 2	** $\beta/2$ **	**1-**β	$\beta/2$
Produit 3	0	β	**1-**β

Ces étapes sont alors appliquées à chaque attribut puis chaque table de codage flou est triée et fusionnée de sorte que les données du premier sujet soient affichées dans les p premières colonnes, les données du deuxième sujet sont affichées dans les p colonnes suivantes, et ainsi de suite, p étant le nombre d'attributs multiplié par trois.

Voici un exemple de tableau qui peut être obtenu avec deux attributs, après ce prétraitement :

	Sujet 1					
	Attribut 1			Attribut 2		
	Pas assez	**JAR**	**Trop**	**Pas assez**	**JAR**	**Trop**
Produit 1	0.05	0.9	0.05	0.9	0.1	0
Produit 2	0	0.1	0.9	0.05	0.9	0.05
Produit 3	0.05	0.9	0.05	0	0.1	0.9

Enfin, XLSTAT utilisera ce tableau pour faire une analyse CATATIS ou CLUSCATA classique.

Boîte de dialogue

Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.

OK : cliquez sur ce bouton pour lancer les calculs.

Annuler : cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

Aide : cliquez sur ce bouton pour afficher l'aide.

 : cliquez sur ce bouton pour rétablir les options par défaut.

 : cliquez sur ce bouton pour effacer les sélections de données.

 : cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT vous permet de sélectionner les données par colonnes ou par plage. Si la flèche est vers la droite, XLSTAT vous permet de sélectionner les données par lignes ou par plage.

 : cliquez sur ce bouton pour changer la façon dont vous chargez les données dans XLSTAT. Si le bouton possède une icône de souris, XLSTAT vous permet de faire une sélection à la souris des données. Si le bouton possède une icône de liste, XLSTAT vous permet de sélectionner les données à partir d'une liste. Si le bouton possède une icône orange, des boutons complémentaires avec un point d'interrogation  sont affichés pour vous permettre d'importer les données à partir de fichiers.

Onglet **Général** :

Données sur l'échelle JAR : sélectionnez les données mesurées sur l'échelle JAR. Plusieurs colonnes peuvent être sélectionnées. Si des en-têtes de colonnes ont été sélectionnés, veuillez vérifier que l'option « Libellés des colonnes » est activée.

Echelle : choisissez le type d'échelle (valeur par défaut : 1 -> 5).

Produits : sélectionnez la colonne contenant les identifiants des produits. Si des en-têtes de colonnes ont été sélectionnés, veuillez vérifier que l'option « Libellés des colonnes » est activée.

Sujets : sélectionnez la colonne contenant les identifiants des sujets. Si des en-têtes de colonnes ont été sélectionnés, veuillez vérifier que l'option « Libellés des colonnes » est activée.

Remarque : il est important que ***tous les sujets aient vu tous les produits*** *et chaque* ***combinaison sujet/produit doit exister et n'exister qu'une seule fois****.*

Plage : activez cette option pour que les résultats soient affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des colonnes : activez cette option si la première ligne des données sélectionnées contient un libellé.

Analyse : choisissez entre une des deux analyses suivantes :

- **CATATIS** : choisissez cette option pour étudier et visualiser les liens entres les produits et les attributs et étudier les accords entre les sujets en utilisant la méthode CATATIS.
- **CLUSCATA** : choisissez cette option pour constituer des classes homogènes de sujets sur la base de leurs perceptions des produits en utilisant la méthode CLUSCATA.

Onglet **Options** :

Bêta : indiquez le paramètre d'accord entre JAR et les autres réponses. Veuillez entrer une valeur entre 0 et 0.5 (valeur par défaut : 0.1).

Filtrer les facteurs : vous pouvez activer l'une des deux options suivantes afin de réduire le nombre de facteurs pour lesquels les résultats sont affichés :

- **% minimum** : activez cette option puis saisissez le pourcentage minimum de la variabilité totale que doivent représenter les facteurs retenus.
- **Nombre maximum** : activez cette option pour fixer le nombre maximum de facteurs à prendre en compte.

Options propres à CLUSCATA :

Troncature : activez cette option si vous voulez que XLSTAT définisse **automatiquement** une troncature, et donc le nombre de classes à retenir, ou si vous voulez définir vous-même le **nombre de classes** à créer, ou si vous voulez définir le **niveau** auquel le dendrogramme doit être tronqué.

Consolidation : activez cette option pour réaliser une consolidation des classes obtenues à partir du dendrogramme.

Classe K+1 : activez cette option pour ajouter une classe supplémentaire qui contiendra les sujets ne se conformant à aucune des classes.

Paramètre rho : choisissez la façon dont vous voulez définir le paramètre rho qui représente l'accord minimal pour être considéré comme suffisamment en accord pour être conservé dans

une classe. Plus ce paramètre augmente, plus l'accord demandé avec la classe est fort et plus vous risquez de placer de sujets dans la classe K+1.

- **Automatique** ;
- **Défini par l'utilisateur**.

Onglet **Données manquantes** :

Ne pas accepter les valeurs manquantes : activez cette option pour que XLSTAT empêche la poursuite des calculs si des valeurs manquantes sont détectées.

Supprimer les observations : activez cette option pour supprimer les observations comportant des données manquantes.

Estimer les données manquantes : activez cette option pour estimer les données manquantes avant le début des calculs.

- **Plus proche voisin** : activez cette option pour estimer les données manquantes d'une observation en recherchant le plus proche voisin de l'observation.

Onglet **Sorties** :

Tableau Produits/Sujets : activez cette option pour afficher le tableau Produits/Sujets qui a été construit durant l'analyse.

Valeurs propres de l'AFC : activez cette option pour afficher le tableau des valeurs propres de l'AFC sur le consensus ou sur le consensus de chaque classe dans le cas d'une analyse CLUSCATA.

Coordonnées de l'AFC : activez cette option pour afficher les coordonnées du consensus de l'AFC dans l'espace des facteurs. Dans le cas d'une analyse CLUSCATA, ce sont les coordonnées du consensus de chaque classe qui sont affichées.

Matrice de similarité (S) : activez cette option pour afficher la matrice des indices de similarité (cosinus de Salton).

Configuration consensus : activez cette option pour afficher la configuration consensus ou la configuration consensus de chaque classe créée par CATATIS.

Similarité sujets/consensus : activez cette option pour afficher le coefficient de similarité entre chaque sujet et le consensus. Dans le cas d'une analyse CLUSCATA il s'agit du coefficient de similarité entre chaque sujet et le consensus de chaque classe.

Poids : activez cette option pour afficher les poids des sujets créés et utilisés par CATATIS.

Homogénéités : activez cette option pour afficher l'homogénéité des sujets. Dans le cas d'une analyse CLUSCATA, il s'agit de l'homogénéité des sujets de chacune des classes ainsi que l'homogénéité globale.

Erreur globale : activez cette option pour afficher l'erreur du critère de minimisation CLUSCATA, équivalente à la variance intra-classes, ou pour afficher l'erreur du critère CATATIS.

Sous-onglet **CLUSCATA** :

Statistiques des nœuds : activez cette option pour afficher les statistiques des nœuds du dendrogramme.

Composition des classes : activez cette option pour afficher la composition de chacune des classes.

Graphiques :

Valeurs propres de l'AFC : activez cette option pour afficher le graphique (*scree plot*) des valeurs propres de l'AFC sur le consensus ou sur le consensus de chaque classe dans le cas d'une analyse CLUSCATA.

Biplot de l'AFC : activez cette option pour afficher le graphique des coordonnées du consensus. Dans le cas d'une analyse CLUSCATA, le graphique des cordonnées du consensus de chaque classe dans l'espace des facteurs sera affiché.

Graphiques sur deux axes : activez cette option si vous souhaitez que les différentes représentations graphiques ne soient affichées que sur les deux premiers axes.

Poids : activez cette option pour afficher le diagramme en bâtons des poids créés et utilisés par CATATIS.

Similarité sujets/consensus : activez cette option pour afficher le diagramme en bâtons des coefficients de similarité entre chaque sujet et le consensus. Dans l'analyse CLUSCATA, il s'agira des coefficients de similarité entre chaque sujet et le consensus de sa classe.

Sous-onglet **CLUSCATA** :

Diagramme des niveaux : activez cette option pour afficher le diagramme des niveaux permettant d'observer l'impact des regroupements successifs sur la variance intra-classes.

Dendrogramme : activez cette option pour afficher le dendrogramme.

- **Horizontal** : choisissez cette option pour afficher un dendrogramme horizontal.
- **Vertical** : choisissez cette option pour afficher un dendrogramme vertical.
- **Tronqué** : activez cette option pour afficher le dendrogramme tronqué (le dendrogramme commence au niveau de la troncature).
- **Etiquettes** : activez cette option pour afficher les libellés des sujets (dendrogramme complet) ou des classes (dendrogramme tronqué) sur le dendrogramme.
- **Couleurs** : activez cette option pour utiliser des couleurs pour représenter les différents groupes sur le dendrogramme complet.

Résultats

Le **tableau Produits/Sujets** est affiché et présente les éléments suivants :

- la première ligne affiche les libellés des descripteurs, qui correspondent aux mêmes libellés entrés dans le champ **Données sur l'échelle JAR** suivis du terme "pasassez", "JAR", ou "trop" ;

- la première colonne contient les libellés des **produits** ;
- enfin, le corps du tableau comprend les valeurs floues (voir la section description).

Valeurs propres de l'AFC : les valeurs propres de l'AFC et le graphique (*scree plot*) correspondant sont affichés.

Coordonnées des produits : les coordonnées des produits du consensus dans l'espace des facteurs sont affichées, ainsi que les graphiques correspondants (en fonction du nombre de facteurs choisi).

Coordonnées des attributs : les coordonnées des attributs du consensus dans l'espace des facteurs sont affichées, ainsi que les graphiques correspondants (en fonction du nombre de facteurs choisi).

Matrice de similarité (S) : la matrice des coefficients de similarité entre tous les sujets est affichée. Le coefficient de similarité utilisé est le cosinus de Salton qui est compris entre 0 et 1. Plus il est proche de 1, plus la similarité est forte. Cette matrice est utilisée par CATATIS pour calculer les poids des sujets.

Poids de chaque sujet : les poids calculés par CATATIS sont affichés, ainsi que le diagramme en bâtons associé. Plus un poids est grand, plus le sujet a contribué à l'élaboration du consensus. Sachant que CATATIS donne du poids aux sujets les plus proches du point de vue global, un poids beaucoup plus faible que les autres signifiera que le sujet est atypique.

Configuration consensus : la configuration consensus créée par CATATIS est affichée. Elle correspond à la moyenne pondérée des données initiales par les poids de CATATIS.

Homogénéité : l'homogénéité des sujets est affichée. C'est une valeur comprise entre $1/m$ (m étant le nombre de sujets) et 1, qui croît avec l'homogénéité des sujets.

Similarité entre chaque sujet et le consensus : les coefficients de similarité entre les sujets et le consensus sont affichés, ainsi que le diagramme en bâtons associé. Tout comme les poids de CATATIS, ces coefficients permettent de détecter des sujets atypiques. L'avantage de ces coefficients est qu'ils sont compris entre 0 et 1, donc plus faciles à interpréter que les poids.

Erreur globale : l'erreur du critère CATATIS est affichée. Elle correspond à la somme de tous les résidus (qui peuvent être présentés par sujet ou par produit).

Résultats propres à CLUSCATA :

Statistiques des nœuds : dans ce tableau sont affichées les informations concernant les nœuds successifs du dendrogramme. Le premier nœud a pour indice le nombre de sujets augmenté de 1. Ainsi, il est aisé de repérer à quel moment un sujet ou un groupe de sujets est regroupé avec un autre groupe de sujets dans le dendrogramme.

Diagramme des niveaux : dans ce graphique sont affichés les niveaux des nœuds du dendrogramme, qui correspondent à l'augmentation du critère de minimisation de CLUSCATA (équivalent à l'augmentation de la variance intra-classes) lors de la fusion de deux classes.

Dendrogrammes : le dendrogramme complet permet de visualiser le regroupement progressif des sujets. Si une troncature a été demandée, un trait en pointillé marque le niveau auquel est effectuée la troncature. Le dendrogramme tronqué permet de visualiser les classes après la troncature.

Exemple

Un tutoriel pour analyser des données JAR est disponible sur le Centre d'aide XLSTAT :

http://www.xlstat.com/demo-jar.htm

Bibliographie

Llobell, F., Vigneau, E. & Qannari, E. M. (September 14, 2022). Multivariate data analysis and clustering of subjects in a Just about right task. *Eurosense*, Turku, Finland.

Llobell, F., Cariou, V., Vigneau, E., Labenne, A., Qannari, E. M. (2019). A new approach for the analysis of data and the clustering of subjects in a CATA experiment. *Food Quality and Preference*, 72, 31-39.

Llobell, F., Giacalone, D., Labenne, A., Qannari, E.M. (2019). Assessment of the agreement and cluster analysis of the respondents in a CATA experiment. *Food Quality and Preference*, 77, 184-190.

Analyse de données RATA

Utilisez l'analyse de données RATA pour analyser vos données Rate-All-That-Apply (RATA). Cette méthode permet :

- d'étudier et visualiser les liens entres les produits et les attributs;
- d'étudier les accords entre les sujets.

Dans cette section :

Description

L'épreuve RATA est une méthode utilisée en analyse sensorielle pour collecter et analyser des données sur la perception des produits par les consommateurs.

La méthode RATA est utile pour évaluer les caractéristiques sensorielles des produits, afin d'éclairer le développement des produits et le contrôle de la qualité dans des secteurs tels que l'alimentation et les boissons, les cosmétiques et les biens de consommation.

Les participants évaluent les attributs du produit, qui peuvent porter sur le goût, l'arôme, la texture ou l'apparence, à l'aide d'une échelle numérique ou d'un système d'évaluation. Les évaluations sont analysées pour identifier les tendances à l'aide d'une analyse statistique multivariée.

Dans l'analyse de données RATA, le fait que les sujets aient eu plusieurs sessions est autorisé. Dans ce cas, l'utilisateur choisira l'option qu'il préfère parmi la moyenne par sujet (si pour un produit et un attribut donné, il a noté une fois 2 et l'autre fois 0, sa moyenne sera de 1) ou bien d'accorder un prédominance à la valeur 0 (si pour un produit et un attribut donné, il a noté une fois 2 et l'autre fois 0, c'est la valeur 0 qui sera retenue, ce qui implique qu'un sujet indiquant au moins une fois qu'un attribut n'est pas présent considère définitivement que cet attribut n'est pas présent).

De plus, des tests de consistance du panel globalement et par attribut sont proposés afin de déterminer si certains attributs ne sont pas compris ou donnent lieu à des réponses trop divergentes. Des tests sur les poids pour déterminer si certains sujets ont un poids non significatif peuvent également être réalisés. Cette dernière option est particulièrement utile lorsque nous avons affaire à des experts.

L'analyse de données RATA est une méthode qui peut se décomposer en 2 grandes parties : * la réalisation d'une ANOVA pour chacun des attributs de manière à vérifier si l'effet du produit est significatif;

- l'utlisation de CATATIS pour éclairer les liens entres attributs et produits.

Utilisation de CATATIS

Il existe plusieurs applications pour CATATIS, parmi lesquelles :

- l'étude et la visualisation des produits et attributs dans les plans factoriels principaux;
- l'étude de la similarité entre les sujets, notamment pour trouver les plus atypiques.

Principe de CATATIS

L'objectif de CATATIS est de former une configuration consensus qui reflète au mieux les différents sujets. Ce consensus peut ensuite être projeté sur différents axes foctoriels à l'aide d'une Analyse Factorielle des correspondances (AFC) ou bien d'une Analyse en Composantes Principales (ACP). Si l'information associée aux 2 ou 3 premiers axes représente un pourcentage suffisant de la variabilité totale du consensus, on pourra représenter les produits et les attributs sur un graphique à 2 ou 3 dimensions, facilitant ainsi grandement l'interprétation.

Interprétation des résultats

La représentation des produits et attributs dans l'espace des k facteurs permet d'interpréter visuellement les proximités entre les produits et les attributs, moyennant certaines précautions.

On peut considérer que la projection d'un produit ou d'un attribut sur un plan est fiable si elle est éloignée du centre du graphique.

Nombre de facteurs

Deux méthodes sont communément utilisées pour déterminer le nombre de facteurs à retenir pour l'interprétation des résultats :

- Regarder la courbe décroissante des valeurs propres. Le nombre de facteurs à retenir correspond au premier point d'inflexion sur la courbe.
- On peut aussi se baser sur le pourcentage cumulé de variabilité représenté par les axes factoriels et décider de se contenter d'un certain pourcentage.

Représentations graphiques

Les représentations graphiques ne sont fiables que si la somme des pourcentages de variabilité associés aux axes de l'espace de représentation, est suffisamment élevée. Si ce pourcentage est élevé (par exemple 80%), on peut considérer que la représentation est fiable. Si le

pourcentage est faible, il est conseillé de faire des représentations sur plusieurs paires d'axes afin de valider l'interprétation faite sur les deux premiers axes factoriels.

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.

OK : cliquez sur ce bouton pour lancer les calculs.

Annuler : cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

Aide : cliquez sur ce bouton pour afficher l'aide.

 : cliquez sur ce bouton pour rétablir les options par défaut.

 : cliquez sur ce bouton pour effacer les sélections de données.

 : cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.

 : cliquez sur ce bouton pour changer la façon dont vous chargez les données dans XLSTAT. Si le bouton possède une icône de souris, XLSTAT vous permet de faire une sélection à la souris des données. Si le bouton possède une icône de liste, XLSTAT vous permet de sélectionner les données à partir d'une liste. Si le bouton possède une icône orange, des boutons complémentaires avec un point d'interrogation  sont affichés pour vous permettre d'importer les données à partir de fichiers.

Onglet **Général**:

Données RATA : sélectionnez les données correspondant aux différents sujets. Si la première ligne de la sélection comprend des en-têtes, l'option « Libellés des attributs » doit être activée.

Libellés des produits : sélectionnez les produits correpondants aux données RATA. Si l'option « Libellés des attributs » est activée, la première cellule de la sélection doit comprendre un en-tête.

Libellés des sujets : sélectionnez les sujets correpondants aux données RATA. Si l'option « Libellés des attributs » est activée, la première cellule de la sélection doit comprendre un en-tête.

Sessions : activez cette option si vous voulez utiliser des sessions dans l'analyse. Si l'option « Libellés des attributs » est activée, la première cellule de la sélection doit comprendre un en-tête.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des attributs : activez cette option si la première ligne des données sélectionnées (Données RATA, Libellés des produits, Libellés des sujets, Sessions) contient un libellé. Si vous n'activez pas cette option, des libellés seront automatiquement créés.

Onglet **Options**:

Filtrer les facteurs : vous pouvez activer l'une des deux options suivantes afin de réduire le nombre de facteurs pour lesquels les résultats sont affichés :

- **% minimum** : activez cette option puis saisissez le pourcentage minimum de la variabilité totale que doivent représenter les facteurs retenus.
- **Nombre maximum** : activez cette option pour fixer le nombre maximum de facteurs à prendre en compte.

Intervalle de confiance (%) : entrez l'intervalle de confiance pour les tests.

Nombre de permutations : entrez le nombre de permutations à réaliser pour les tests.

Prétraitement des sessions : choisissez la méthode à utiliser pour effectuer le prétraitement des sessions.

Moyenne : activez cette option si vous souhaitez utiliser la moyenne.

Dominance de la valeur 0 : activez cette option si vous souhaitez que la valeur 0 soit considérée comme dominante.

Onglet **Données manquantes** :

Ne pas accepter les valeurs manquantes : activez cette option pour que XLSTAT empêche la poursuite des calculs si des valeurs manquantes sont détectées.

Remplacer les valeurs manquantes par 0 : activez cette option si vous considérez que les valeurs manquantes sont équivalentes à des 0.

Onglet **Sorties**:

Matrice de similarité (S) : activez cette option pour afficher la matrice des indices de similarité (Cosinus de Salton).

Facteurs de mise à l'échelle : activez cette option pour afficher les facteurs de mise à l'échelle des sujets.

Poids : activez cette option pour afficher les poids créés et utilisés par CATATIS.

Tests des poids : activez cette option pour tester si les poids des sujets sont significatifs.

Configuration consensus : activez cette option pour afficher la configuration consensus créée par CATATIS.

Similarité sujets/consensus : activez cette option pour afficher le coefficient de similarité entre chaque sujet et le consensus.

Homogénéité : activez cette option pour afficher l'homogénéité des sujets.

Tests de consistance : activez cette option pour tester si le panel est consistant globalement et par attribut.

Résidu par sujet : activez cette option pour afficher les résidus de CATATIS pour chaque sujet.

Résidu par produit : activez cette option pour afficher les résidus de CATATIS pour chaque produit.

Erreur globale : activez cette option pour afficher l'erreur du critère CATATIS.

Réduction dimension : choisissez la méthode (AFC ou ACP) pour projeter la configuration consensus.

Valeurs propres : activez cette option pour afficher le tableau des valeurs propres de l'AFC ou de l'ACP sur le consensus.

Coordonnées : activez cette option pour afficher les coordonnées du consensus dans l'espace des facteurs.

Onglet **Graphiques** :

Valeurs propres : activez cette option pour afficher le graphique (*scree plot*) des valeurs propres de l'AFC ou de l'ACP sur le consensus.

Biplot : activez cette option pour afficher le graphique des coordonnées du consensus dans l'espace des facteurs.

Graphiques sur 2 axes : activez cette option pour que XLSTAT ne vous demande pas de sélectionner les axes, et affiche automatiquement les 2 premiers.

Facteurs de mise à l'échelle : activez cette option pour afficher le diagramme en bâtons des facteurs de mise à l'échelle des sujets.

Poids : activez cette option pour afficher le diagramme en bâtons des poids créés et utilisés par CATATIS.

Similarité sujets/consensus : activez cette option pour afficher le diagramme en bâtons des coefficients de similarité entre chaque sujet et le consensus.

Graphiques de corrélations : activez cette option pour afficher le graphique mettant en jeu les corrélations entre les composantes et les variables initiales. Ce graphique est communément appelé cercle de corrélation.

Résidu par sujet : activez cette option pour afficher le diagramme en bâtons des résidus du critère CATATIS pour chaque sujet.

Résidu par produit : activez cette option pour afficher le diagramme en bâtons des résidus du critère CATATIS pour chaque produit.

Résultats

Résumés des ANOVA : le résumé des ANOVA pour chacun des attributs est affiché.

Répétabilité des sujets : Le coefficient de similarité (Cosinus de Salton) entre les résultats des différentes sessions est affiché. Ce coefficient prend des valeurs entre 0 et 1 et croît avec la ressemblance entre les sessions.

Valeurs propres : les valeurs propres de l'AFC ou de l'ACP et le graphique (*scree plot*) correspondant sont affichés.

Coordonnées des produits : les coordonnées des produits du consensus dans l'espace des facteurs sont affichées, ainsi que les graphiques correspondants (en fonction du nombre de facteurs choisi).

Coordonnées des attributs : les coordonnées des attributs du consensus dans l'espace des facteurs sont affichées, ainsi que les graphiques correspondants (en fonction du nombre de facteurs choisi).

Matrice de similarité (S) : la matrice des coefficients de similarité entre tous les sujets est affichée. Le coefficient de similarité utilisé est le cosinus de Salton. Il est compris entre 0 et 1. Plus il est proche de 1, plus la similarité est forte. Cette matrice est utilisée par CATATIS pour calculer les poids des sujets.

Facteurs d'échelle pour chaque sujet : les facteurs d'échelle sont affichés, ainsi que le diagramme en bâtons associé. Plus un facteur d'échelle d'un sujet est grand, plus le barème original du sujet est faible.

Poids de chaque sujet : les poids calculés par CATATIS sont affichés, ainsi que le diagramme en bâtons associé. Plus un poids est grand, plus le sujet a contribué à l'élaboration du consensus. Sachant que CATATIS donne du poids aux sujets les plus proches du point de vue global, un poids beaucoup plus faible que les autres signifiera que le sujet est atypique.

Tests des poids : les résultats des tests des poids sont affichés. Si un sujet a son poids non significatif, alors son point de vue est très différent du point de vue global, et ses résultats peuvent être remis en cause s'il s'agit d'un expert.

Configuration consensus : la configuration consensus créée par CATATIS est affichée. Elle correspond à la moyenne pondérée des données initiales par les poids de CATATIS.

Homogénéité : l'homogénéité des sujets est affichée. C'est une valeur comprise entre $1/m$ (m étant le nombre de sujets) et 1, qui croît avec l'homogénéité des sujets.

Tests de consistance : les résultats des tests de consistance sont affichés globalement et par attribut. Si le panel est globalement non consistant, les données peuvent malheureusement être jetées. Si il est non consistant pour un ou plusieurs attributs, alors ces attributs sont sujets à tellement de divergence qu'ils ont sûrement été mal compris.

Distance entre la médiane des permutations et l'homogénéité: cette distance permet de voir à quel point l'homogénéité des sujets est forte par rapport à des réponses aléatoires.

Similarité entre chaque sujet et le consensus : les coefficients de similarité entre les sujets et le consensus sont affichés, ainsi que le diagramme en bâtons associé. Tout comme les poids de CATATIS, ces coefficients permettent de détecter des sujets atypiques. L'avantage de ces coefficients est qu'ils sont compris entre 0 et 1, donc plus faciles à interpréter que les poids.

Erreur globale : l'erreur du critère CATATIS est affichée. Elle correspond à la somme de tous les résidus (qui peuvent être présentés par sujet ou par produit).

Résidu par sujet : ce tableau et le diagramme en bâtons correspondant permettent de visualiser la répartition des résidus de CATATIS par sujet. On peut ainsi repérer pour quels sujets CATATIS a été moins efficace, autrement dit, quels sujets se démarquent le plus de la configuration consensus.

Résidu par produit : ce tableau et le diagramme en bâtons correspondant permettent de visualiser la répartition des résidus de CATATIS par produit. On peut ainsi repérer pour quels produits CATATIS a été moins efficace, autrement dit, quels produits se démarquent le plus de la configuration consensus.

Exemple

Un tutoriel sur la façon d'utiliser la fonctionnalité RATA est disponible sur le Help Center de XLSTAT:

http://www.xlstat.com/demo-rataf.htm

Bibliographie

Bonnet, L., Ferney, T., Riedel, T., Qannari, E.M., Llobell, F. (September 14, 2022). Using CATA for sensory profiling: assessment of the panel performance. Eurosense, Turku, Finland.

Bonnet, L., Llobell, F., Qannari, E.M. (Pangborn 2023). Assessment of the panel performance in a RATA experiment.

Llobell, F., Cariou, V., Vigneau, E., Labenne, A., Qannari, E. M. (2019). A new approach for the analysis of data and the clustering of subjects in a CATA experiment. *Food Quality and Preference*, **72**, 31-39.

Llobell, F., Giacalone, D., Labenne, A., & Qannari, E. M. (2019). Assessment of the agreement and cluster analysis of the respondents in a CATA experiment. *Food Quality and Preference*, **77**, 184-190.

Llobell, F. (2020). Classification de tableaux de données, applications en analyse sensorielle (Doctoral dissertation, Nantes, Ecole nationale vétérinaire).

Outils pour le marketing

Taille d'échantillon

Utilisez cet outil pour calculer le nombre de répondants nécessaires afin d'obtenir des résultats statistiquement solides pour une certaine population ou d'obtenir la marge d'erreur de votre échantillon.

Dans cette section :

Description

Lorsque l'on parle d'étude statistique ou d'étude de marché impliquant des recherches sur des groupes de populations, il devient important de se poser la question de la taille de l'échantillon. Afin d'obtenir des bons résultats il faut en effet avoir un échantillon le plus représentatif possible de la population totale. Dans le but de se raprocher de la représentativité de la population totale, XLSTAT permet, avec cet outil, de calculer le bon nombre de personnes à interroger pour obtenir un échantillon ni trop grand (ce qui rendrait l'étude plus complexe et couteuse), ni trop petit (ce qui engendrerait des résultats éronnés). Si vous connaissez déjà la taille de votre échantillon, XLSTAT permet aussi de vérifier la marge d'erreur de celui ci.

La formule suivante est utilisée pour calculer les résultats :

$$\text{Taille de l'échantillon} = \frac{Z^2 * p * (1-p)}{(\text{marge d'erreur})^2}$$

où Z est le score de la loi normale défini à partir de l'intervalle de confiance rentré dans l'interface, p est la proportion de la population présentant la caractéristique étudiée (par soucis de simplicité, elle est automatiquement choisie à 0.5 dans XLSTAT), et la marge d'erreur est la différence que vous acceptez entre la moyenne de l'échantillon et la moyenne de la population.

Boîte de dialogue

OK : cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue.

: cliquez sur ce bouton pour afficher l'aide.

 : cliquez sur ce bouton pour rétablir les options par défaut.

Objectif : choisissez entre le calcul de la taille d'échantillon et le calcul de la marge d'erreur (en fonction de ce choix les champs suivants seront différents).

Taille de la population : entrez la taille de la population étudiée.

Marge d'erreur (dans le cas du calcul de la taille d'échantillon) : entrez la marge d'erreur que vous acceptez pour votre étude (en %).

Nombre de répondants (dans le cas du calcul de la marge d'erreur) : entrez le nombre de personnes ayant répondu à votre étude.

Intervalle de confiance : entrez la taille de l'intervalle de confiance désiré (en %).

Taux de réponse estimé : sélectionnez cette option si vous souhaitez calculer le nombre d'invitations requises pour atteindre la bonne taille d'échantillon. Entrez alors le taux de réponse estimé (en %).

Calculer : cliquez sur ce bouton pour afficher le résultat dans la zone *Résultats* de la boîte de dialogue.

Afficher les résultats dans :

- **Plage** : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.
- **Feuille** : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.
- **Classeur** : activez cette option pour afficher les résultats dans un nouveau classeur.

Résultats : les résultats sont affichés dans cette zone.

Effacer : cliquez sur ce bouton pour effacer les résultats enregistrés dans la zone *Résultats* de la boîte de dialogue.

Résultats

Les résultats affichés par XLSTAT sont un tableau récapitulatif avec la taille de la population, la marge d'erreur, l'intervalle de confiance, la taille d'échantillon et le nombre d'invitations requises.

Waffle Chart : permet de représenter le pourcentage de la taille de l'échantillon comparé à la taille de la population.

Diagramme en secteurs : permet de représenter la taille de l'échantillon par rapport au nombre d'invitations requises.

Exemple

Un exemple d'utilisation du calculateur de taille d'échantillon est disponible sur le Centre d'aide XLSTAT :

https://www.xlstat.com/demo/samplesize_fr

Price Sensitivity Meter (Van Westendorp)

Utilisez cet outil pour évaluer la gamme de prix adéquate pour un produit donné, ou le prix optimal pour le volume ou le chiffre d'affaires, sur la base de données collectées auprès de consommateurs.

Dans cette section :

Description

Cette version de l'analyse de la sensibilité au prix (du volume de vente et du chiffre d'affaires) est due à Van Westendorp qui l'a présentée à l'occasion du congrès de l'association ESOMAR en 1976. Elle a par la suite été enrichie par Newton *et al.* en 1993, afin de palier le reproche régulier qui était fait à cette méthode, de ne pas prendre en compte l'intention d'achat.

Cette méthode consiste en une enquête menée auprès d'un panel de consommateurs, auxquels on demande, pour un produit donné, quels sont les prix seuils auxquels ils estiment que le produit est : + *trop bon marché* (au point que cela discrédite le produit, noté TBM) + *bon marché* (un prix auquel ils estiment que cela le rend attractif, noté BM) + *cher* (le prix auquel ils estiment que le produit est dans la fourchette haute de ce qu'ils attendraient, CH) + *trop cher* (un prix rédhibitoire, TCH)

Le résultat caractéristique de ce type d'analyse de sensibilité au prix est un graphique présentant une série de courbes dont les intersections déterminent des prix critiques : à partir des données de l'enquête on construit six courbes de distributions cumulées (ou leurs opposées), d'abord pour les quatre types de prix recueillis (*trop bon marché*, *bon marché*, *cher*, *trop cher*), puis, par déduction, on calcule la distribution pour *pas bon marché* et *pas cher*.

Pour *trop bon marché* et *bon marché*, on prend l'opposé de la courbe de distribution. On calcule donc pour chacun des prix indiqués par les consommateurs, quelle proportion de consommateurs a indiqué un prix supérieur. Ces courbes diminuent donc de 1 à 0 lorsque le prix augmente. Pour *cher* et *trop cher*, on prend la courbe de distribution cumulée (aussi appelée fonction de répartition empirique). On calcule donc pour chacun des prix indiqués par les consommateurs, quelle proportion de consommateurs a indiqué un prix inférieur. Ces courbes augmentent donc de 0 à 1 lorsque le prix augmente.

L'intersection entre les courbes *bon marché* et *cher* correspond au prix pour lequel autant de personnes considèrent que le produit est cher que bon marché. Même s'il n'est pas forcément fort, il y a donc un désaccord entre deux groupes de même taille de consommateurs sur ce prix qui a été dénommé **Indifference Price (IDP)**. D'après Van Westendorp, ce prix correspond à une réalité du marché. L'IDP peut être interprété comme le prix médian sur le marché de ce

type de produits ou comme le prix proposé par un leader du marché. En principe, pour une large majorité de consommateurs, ce prix est bon, un faible effectif le trouvant *bon marché* (donc probablement pas assez cher pour l'entreprise qui le commercialise) ou *cher* (donc potentiellement à risque vis à vis d'un concurrent) et la majorité le trouvant très probablement correct. Bien entendu, un effectif encore plus faible le trouvera *trop bon marché* (qualité suspecte) ou *trop cher* (inaccessible).

L'intersection entre les courbes des *trop bon marché* et des *trop cher* correspond au prix pour lequel autant de personnes considèrent que le produit est *trop cher* ou *trop bon marché*. En principe ce prix qui correspond à des avis totalement opposés concerne peu de consommateurs. Ce prix est nommé le **Optimal Pricing Point (OPP)**, dans le sens où, en principe, un maximum de consommateurs devrait pouvoir acheter le produit.

La gamme de prix acceptable est donnée par, pour la borne inférieure, l'intersection entre la courbe des prix *trop bon marché* et *pas bon marché* et pour la borne supérieure, par l'intersection entre la courbe des prix *trop cher* et *pas bon marché*. Entre ces deux prix marginaux, les auteurs estiment que les volumes de vente sont élevés.

Intention d'achat

XLSTAT permet de prendre en compte les apports de Newton *et al.* (1993) qui ont proposé de prendre en compte les intentions d'achat dans le cadre de l'enquête, en demandant quel est le niveau d'intention d'achat pour le prix *bon marché* et pour le prix *cher*. Ces scores peuvent être traduits en probabilités, soit automatiquement, soit grâce à un tableau de conversion. Une fois que l'on dispose des probabilités, on peut identifier quel prix est susceptible d'engendrer un nombre de ventes maximal et quel prix est susceptible d'engendrer un chiffre d'affaires maximal.

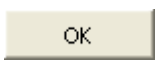
La prise en compte des intentions d'achat permet donc aux responsables en charge de la définition des prix de définir le prix du produit en maximisant les chances que l'impact soit favorable sur le volume ou le chiffre d'affaires en fonction de la stratégie choisie.

Format des données

XLSTAT permet de produire des résultats partiels si seuls les prix *bon marché* et *cher* sont disponibles, mais l'analyse sera moins riche dans ce cas.

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.

 : cliquez sur ce bouton pour rétablir les options par défaut.

 : cliquez sur ce bouton pour effacer les sélections de données.

 : cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.

Onglet **Général**:

Données de prix (TBM/BM/CH/TCH) : sélectionnez les données associées aux prix. Vous devez soit sélectionner deux colonnes (BM/CH) soit quatre colonnes (TBM/BM/CH/TCH). Si des en-têtes de colonnes ont été sélectionnés, veuillez vérifier que l'option « Libellés des variables » est activée.

Vérifier la cohérence : activez cette option si vous voulez que XLSTAT vérifie que les prix sont bien en ordre croissant pour chaque individu. Dans le cas contraire, l'individu n'est pas pris en compte pour l'analyse.

Groupes : activez cette option si vous vous voulez effectuer des analyses par groupe, puis sélectionnez les données indiquant à quel groupe appartient chaque individu.

Intention d'achat : activez cette option si vous disposez des scores d'intention d'achat pour les prix *bon marché* et *cher*. Si des en-têtes de colonnes ont été sélectionnés, veuillez vérifier que l'option « Libellés des variables » est activée. Si cette option est activée et que les intentions d'achat ne sont pas des probabilités et que le tableau de conversion n'est pas fourni, XLSTAT fait automatiquement la conversion.

Tableau de conversion : activez cette option si vous disposez d'un tableau permettant de convertir les scores en probabilités.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des variables : activez cette option si la première ligne des données sélectionnées contient un libellé.

Libellés des observations : activez cette option si vous voulez utiliser des libellés pour les observations pour l'affichage des résultats. Si l'option « Libellés des variables » est activée, la première cellule de la sélection doit comprendre un en-tête. Si vous n'activez pas cette option, des libellés seront automatiquement créés (Obs1, Obs2, …).

Poids : activez cette option si vous voulez pondérer les observations. Si vous n'activez pas cette option, les poids seront tous considérés comme valant 1. Les poids doivent être impérativement supérieurs ou égaux à 0. Si un en- tête de colonne a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

Onglet **Options**:

Ajuster des lois normales : activez cette option si vous voulez ajuster des lois normales aux différents échantillons de prix et utiliser les lois ajustées pour calculer les différents prix et intervalles de prix.

Onglet **Sorties**:

Résultats individuels : activez cette option pour afficher les probabilités d'achat pour chaque individu au prix donnant le volume le plus élevé et au prix donnant le chiffre d'affaire le plus élevé.

Résultats

Statistiques descriptives : le tableau de statistiques descriptives présente pour les différents prix ainsi que l'amplitude, le nombre d'observations, le nombre de données manquantes, le nombre de données non manquantes, la moyenne, et l'écart-type (non biaisé).

Box plots : des box plots sont affichés afin de visualiser rapidement la distribution des différents prix. Si les individus appartiennent à différents groupes déclarés dans la boîte de dialogue, pour chaque type de prix, un graphique comparant les différents groupes est produit.

Statistiques : dans ce tableau sont affichés le **Indifference Price (IDP)** et le **Optimal Pricing Point (OPP)**. Le prix et la proportion d'individus correspondant à ces points sont donnés.

La **gamme de prix acceptable** est ensuite donnée.

Le graphique **Price Sensitivy Meter** est le graphique principal de la méthode qui présente les différentes courbes de prix cumulées. Sur ce graphique sont affichés l'**IDP**, l'**OPP** et la gamme de prix acceptable.

Si les données d'intention d'achat ont été renseignées, le tableau qui suit indique les prix optimaux pour le volume et le chiffre d'affaires, ainsi que pour chacun, la probabilité moyenne d'achat, le volume attendu sur la population de l'étude et le chiffre d'affaires attendu.

Les courbes du volume et du chiffre d'affaire en fonction des prix sont également affichées. Le dernier tableau présente, si l'option est activée dans la boîte de dialogue, la probabilité d'achat pour les prix optimaux pour le volume et le chiffre d'affaires.

Exemple

Un exemple d'utilisation du *Price Sensitivity Meter* est disponible sur le Centre d'aide XLSTAT à l'adresse

http://www.xlstat.com/demo-psmeterf.htm

Bibliographie

Newton D., Miller J. and Smith P. (1993). A market acceptance extension to traditional price sensitivity measurement. *Proceedings of the American Marketing Association Advanced Research Techniques Forum*.

Van Westendorp P. (1976). NSS-Price Sensitivity Meter (PSM) – A new approach to study consumer perception of price. *Proceedings of the 29th ESOMAR Congress*, 139-167.

Elasticité prix de la demande

Utilisez cet outil pour évaluer l'élasticité prix d'un produit et pour déterminer le prix générant le chiffre d'affaires optimal.

Dans cette section :

Description

L'étude de l'**élasticité prix** (Price elasticity of demand, PED ou E_d en anglais) est essentielle en marketing car elle est l'une des approches qui permet de fixer le prix d'un produit, sachant qu'en général, on cherchera, au moins à moyen terme, à maximiser le chiffre d'affaires associé à la vente du produit.

L'élasticité, dont on doit le concept à Alfred Marshall (1920) est définie comme la variation relative de la demande (ou de la quantité vendue) Q lorsque le prix P change. Cela s'écrit mathématiquement :

$$E_d = \frac{dQ/Q}{dP/P} \tag{1}$$

En général, une augmentation des prix est assortie d'une baisse des quantités vendues (même si de nombreux exemples contraires existent, correspondant à ce qui est parfois appelé un effet de snobisme). L'élasticité est donc souvent une quantité négative, bien que certains auteurs prennent l'opposé.

Comme dQ/dP correspond mathématiquement à une variation infinitésimale en un point donné, dans la pratique, on calcule l'élasticité simple (*point elasticity*) ou l'élasticité arc (*arc elasticity*).

L'**élasticité simple** (*point elasticity* ou élasticité point) est définie comme le rapport de la variation relative $(Q2 - Q1)/Q1$ de la quantité vendue, lorsque l'on monte les prix de $P1$ à $P2$, et de la variation relative des prix.

$$E_d = \frac{(Q2 - Q1)/Q1}{(P2 - P1)/P1} \tag{2}$$

Remarques : + La demande est dite élastique au prix si la demande réagit fortement lorsque l'on augmente le prix de $P1$ à $P2$ ($P1Q1 < P2Q2$). Le terme d'élasticité peut se comprendre

comme le fait que la demande réagit comme si elle était liée par un élastique aux prix ($E_d < -1$). Elle est parfaitement élastique si $E_d = -\infty$. Cela se produira si une très faible variation de prix fait chuter fortement les quantités vendues (ou inversement). + L'élasticité est dite unitaire et vaut -1 si le volume évolue comme le prix. En ce qui concerne le chiffre d'affaires, la baisse de la quantité vendue est exactement compensée par la hausse des prix. Par example, une augmentation des prix de 1% entraîne une baisse des quantités vendues de 1% (ou réciproquement). + La demande est dite inélastique (ou rigide, ou faiblement élastique) si la variation du prix à la hausse a peu d'impact sur les volumes vendus $(-1 < E_d \leq 0)$. On a inélasticité totale si $E_d = 0$: la variation des prix est sans effet sur les quantités vendues. + Il existe des cas où la hausse des prix peut entraîner une augmentation des quantités ($E_d > 0$). Il s'agit de produits de luxe (biens de type Veblen) ou au contraire de produits de nécessité de qualité inférieure (biens de type Giffen) ou de situations particulières (changement de positionnement d'un produit). + On ne peut pas déduire l'évolution du chiffre d'affaires de l'élasticité prix.

Le problème de l'*élasticité simple* est qu'elle n'est pas identique, qu'on la calcule en prenant comme prix de référence $P1$, le prix haut ou le prix bas.

L'**élasticité arc** (*arc elasticity*) est définie comme le rapport de la variation relative de la quantité vendue $Q2$, lorsque l'on monte les prix de $P1$ à $P2$, et de la variation relative des prix.

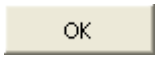
$$E_d = \frac{(Q2 - Q1)/((Q1 + Q2)/2)}{(P2 - P1)/((P1 + P2)/2)} = \frac{(Q2 - Q1)/(Q1 + Q2)}{(P2 - P1)/(P1 + P2)} \quad (3)$$

XLSTAT permet de calculer les deux types d'élasticité.

Il est commun de représenter les courbes reliant quantité et prix en mettant en ordonnée le prix et en abscisse la quantité. Ceci est contre-intuitif dans notre cas, puisque nous sommes ici dans une perspective ou les quantités vendues dépendent des prix, mais s'explique par la théorie économique sous-jacente d'étude de l'offre et de la demande et de l'étude des équilibres de marché, où le prix est déterminé en fonction de l'offre disponible et de la demande. XLSTAT affiche les graphiques suivant les deux alternatives.

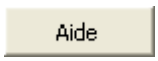
Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.



: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.

 : cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.

Onglet **Général**:

Prix : sélectionnez les données associées aux prix. Si des en-têtes de colonnes ont été sélectionnés, veuillez vérifier que l'option « Libellés des variables » est activée.

Demande : sélectionnez les données correspondant à la demande (quantité vendue) associée à chaque prix. Si des en-têtes de colonnes ont été sélectionnés, veuillez vérifier que l'option « Libellés des variables » est activée.

Groupes : activez cette option si vous voulez effectuer des analyses par groupe, puis sélectionnez les données indiquant à quel groupe appartient chaque individu.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des variables : activez cette option si la première ligne des données sélectionnées (prix, volume et groupe) contient un libellé.

Elasticité simple : activez cette option pour calculer et afficher les élasticités simples.

Elasticité arc : activez cette option pour calculer et afficher les élasticités arc.

Résultats

Statistiques descriptives : le tableau de statistiques descriptives présente pour les prix et la demande, le nombre d'observations, le nombre de données manquantes, le nombre de données non manquantes, la moyenne, et l'écart-type (non biaisé).

XLSTAT affiche dans un premier tableau le prix, la demande, le chiffre d'affaires et l'élasticité simple si elle a été demandée. Si l'élasticité complexe a été demandée, un second tableau affiche pour chaque point médian calculé, le prix, la demande, le chiffre d'affaires et l'élecitcité complexe.

Ces divers élements sont ensuite croisés dans une série de graphiques.

Exemple

Un exemple de calcul de l'elasticité prix de la demande est disponible sur le Centre d'aide XLSTAT à l'adresse

http://www.xlstat.com/demo-elasticityf.htm

Bibliographie

Henderson J. P. (1973). William Whewell's Mathematical Statements of Price Flexibility, Demand Elasticity and the Giffen Paradox. The Manchester School, **41**, 329-342.

Macgregor D. H. (1942). Marshall and His Book. Economica. **9(36)**, 313-324.

Marshall A. (1920). Principles of Economics. *Library of Economics and Liberty*, London.

Customer Lifetime Value (CLV)

A partir des données de vos commandes, la customer lifetime value (CLV) va aider votre entreprise ou votre association à déterminer de combien rapporteront vos clients, et estimer combien de temps vous les conserverez après acquisition. Ceci va vous permettre de rationnaliser les opérations marketing ou publicitaires que vous pourriez engager afin d'augmenter le taux de rétention client.

Dans cette section :

Description

Quels bénéfices attendez-vous de vos clients au cours de leur vie ? Ou encore comment l'augmentation des taux de rétention affecterait les bénéfices futurs de votre entreprise ? Ce sont entre autres les questions auxquelles la CLV pourrait répondre.

Customer Lifetime Value (CLV) : La CLV appelée valeur vie client est un indicateur qui peut être défini comme étant la somme des profits générés par une entreprise tout au long de sa relation avec un client.

Plusieurs modèles de calcul de la CLV existent, ils varient en fonction de plusieurs facteurs qui sont propres à chaque entité. Le modèle implémenté ici est le modèle dit de rétention simple (SRM). Ce modèle est adapté pour les données client en situation contractuelle. Le modèle de rétention simple (SRM) estime la CLV en supposant que :

- le pourcentage de clients retenus chaque période, c'est à dire le taux de rétention, r est constant dans le temps.
- Le cashflow m généré par période n'est pas affecté par le moment d'annulation.
- Le fait qu'un client annule au cours d'une période t est indépendant du fait qu'il annule au cours d'une autre période.

Modèle probabiliste pour le calcul de la CLV :

- **Temps d'annulation/d'attrition** : Supposons que tous les clients d'un segment sont retenus chaque période avec un taux de rétention r et que le fait qu'un client annule au cours d'une période est indépendant du fait qu'il annule au cours d'une autre période. Soit T une variable aléatoire indiquant la période d'annulation (attrition) et t une réalisation de T. Sous ces hypothèses, T a une distribution géométrique. Les probabilités pour une loi géométrique sont données par :

$$f(t) = P(T = t) = r^{t-1}(1 - r)\cdot$$

Cette formule nous donne la probabilité qu'a un client d'annuler au temps t. Elle peut aussi être interprétée comme étant la probabilité qu'a un client de rester $t-1$ périodes.

Aussi XLSTAT propose de calculer les quantiles de T. Le quantile d'ordre α de la variable aléatoire T, apelé P_α divise la distribution de la variable aléatoire T tel que α pourcent de la distribution a $T \leq P_\alpha$ et $1-\alpha$ pourcent de la distribution a $T \geq P_\alpha$. On a donc $P(T \leq P_\alpha) = \alpha$ et $P(T \geq P_\alpha) = 1-\alpha$. Sous les hypothèses du SRM on a :

$$P_\alpha = \frac{\log(1-\alpha)}{\log r}.$$

- **CLV** : lorsqu'un client annule lors de la période t, il y aura eu t cashflows si le règlement a lieu en début de période et $t-1$ cashflows s'il a lieu à la fin. On note d le taux de réduction (discount). Pour une période d'annulation particulière la CLV peut être calculée en utilisant les formules suivantes :

$$CLV = \sum_{t=0}^{T-1} \frac{m}{(1+d)^t} = m \times \frac{(1+d)[1-(1+d)^{-T}]}{d} \qquad \text{réglement: début de période,}$$

$$CLV = \sum_{t=1}^{T} \frac{m}{(1+d)^t} = m \times \frac{1-(1+d)^{-T}}{d} \qquad \text{réglement: fin de période.}$$

Cependant, le moment d'attrition ou de churn T est une variable aléatoire, la CLV a donc une distribution. Les clients ayant de grandes valeurs de T auront une grande CLV. Ainsi nous pouvons résumer la distribution de la CLV par sa moyenne ou espérance.

$$E[CLV] = \frac{m(1+d)}{1+d-r} \qquad \text{réglement: début de période,}$$

$$E[CLV] = \frac{m \times r}{1+d-r} \qquad \text{réglement: fin de période .}$$

- **Estimation des taux de rétention** : Dans la section ci dessus le taux de rétention r était supposé connu, mais en pratique ce n'est pas toujours le cas. C'est dans ce cadre que XLSTAT propose une estimation de ce dernier à partir des données. Parmi les clients d'une entreprise, certains clients (pas tous) ont dèjà annulé ou annuleront leur abonnement. Un client n'ayant pas encore annulé sera dit censuré, l'entreprise n'aura donc pas observé de date d'annulation/d'attrition.

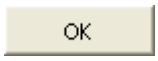
Soit n_0 le nombre de clients n'ayant pas encore annulé et n_1 le nombre de clients ayant déjà annulé leur abonemment. Pour ces derniers nous avons donc une période d'attrition t observée.

Si le client i a déjà annulé, notons t_i la période d'attrition c'est à dire que le client i a été présent pendant $t_i - 1$ périodes. Pour ceux toujours présents notons C_i la période de censure, on a pour ces clients $T_i > C_i$. L'estimation du taux de rétention est faite via la formule suivante :

$$\hat{r} = 1 - \frac{n_1}{\sum_i t_i + \sum_i C_i}.$$

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.


: cliquez sur ce bouton pour changer la façon dont vous chargez les données dans XLSTAT. Si le bouton possède une icône de souris, XLSTAT vous permet de faire une sélection à la souris des données. Si le bouton possède une icône de liste, XLSTAT vous permet de sélectionner les données à partir d'une liste. Si le bouton possède une icône orange, des boutons complémentaires avec un point d'interrogation  sont affichés pour vous permettre d'importer les données à partir de fichiers.

Onglet **Général** :

RMC (Revenu Moyen par Client) : sélectionnez les données associées aux revenus. Vous devez sélectionner une seule colonne. Si des en-têtes de colonnes ont été sélectionnées, veuillez vérifier que l'option « Libellés des variables » est activée.

Temps (acquisition/attrition) : sélectionnez dans cet ordre les deux colonnes : date d'acquisition et date d'attrition. Si un client n'est pas encore parti, alors sa seconde colonne doit être une cellule vide. Si des en-têtes de colonnes ont été sélectionnées, veuillez vérifier que l'option « Libellés des variables » est activée.

Segments : activez cette option si vous voulez effectuer des analyses par segments, puis sélectionnez les données indiquant à quel segment appartient chaque individu. Si des en-têtes

de colonnes ont été sélectionnés, veuillez vérifier que l'option « Libellés des variables » est activée.

Periode de souscription : sélectionnez la période de souscription.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des variables : activez cette option si la première ligne des données sélectionnées contient un libellé.

Noms des clients : activez cette option si vous voulez utiliser des libellés pour les clients pour l'affichage des résultats. Si l'option « Libellés des variables » est activée, la première cellule de la sélection doit comprendre une en-tête. Si vous n'activez pas cette option, des libellés seront automatiquement créés (Obs1, Obs2, …).

Onglet **Options** :

Libellés des segments : sélectionnez les libellés des segments. Ce champ est uniquement disponible lorsque l'option segments est active et lorsqu'on a au moins une des trois options suivante d'activée : coûts fixes, taux de remise ou taux de rétention défini par l'utilisateur.

Taux de remise : activez cette option si vous voulez prendre en compte le taux de remise appliqué aux clients. Ce taux est considéré fixe. Si l'option segments est activée, sélectionnez une valeur par segment et assurez vous que les lignes soient dans le même ordre que les libellés des segments.

Coûts fixes : activez cette option si vous voulez déduire certains coûts fixes d'exploitation des revenus générés lors du calcul de la CLV. Sélectionnez plusieurs colonnes si vous avez plusieurs coûts à inclure. Si l'option segments est activée, sélectionnez une ligne par segment et assurez vous que les lignes soient dans le même ordre que les libellés des segments.

Taux de rétention :

- **Estimer** : activez cette option pour estimer le taux de rétention à partir des données d'entrée. Si l'option segments est activée, le taux de rétention sera estimé pour chaque segment.
- **Défini par l'utilisateur** : activez cette option si vous souhaitez définir vous-même le taux de rétention. Ce taux est supposé fixe. Si l'option segments est activée, sélectionnez une valeur par segment et assurez vous que les lignes soient dans le même ordre que les libellés des segments.

Règlement :

- **Début de période** : sélectionnez cette option si le règlement est effectué en début de période.

- **Fin de période** : sélectionnez cette option si le règlement est effectué en fin de période.

Onglet **Sorties** :

Statistiques descriptives : activez cette option pour afficher les statistiques descriptives de la variable correspondant aux revenus (RMC).

CLV : activez cette option pour afficher la CLV moyenne. Si l'option **segments** est activée, cette valeur est affichée pour chaque segment.

Taux d'attrition estimé : activez cette option pour afficher le taux d'attrition et de rétention. Si l'option **segments** est activée, ces valeurs sont affichées pour chaque segment.

Durée estimée avant défection : activez cette option pour afficher des statistiques concernant la durée avant défection afin de visualiser rapidement la dispersion des temps de départ des clients.

Evolution du cashflow : activez cette option pour afficher l'évolution du cashflow. Chaque ligne de ce tableau contient un prévisionnel du cashflow par période.

Résultats individuels : activez cette option pour afficher la CLV par client.

CLV prévisionnelle : activez cette option si vous voulez faire des simulations sur la CLV.

- **durée** : choisissez la période sur laquelle vous souhaitez faire la simulation.

Analyse de sensibilité : activez cette option pour afficher l'impact d'une augmentation du taux de rétention sur la CLV.

Onglet **Graphiques** :

CLV par segment : activez cette option pour afficher la CLV par segments.

Durée estimée avant défection : activez cette option pour afficher le graphique résumant l'information contenue dans le tableau correspondant décrit plus haut.

Probabilités de churn : activez cette option pour afficher la courbe représentant les probabilités de churn en fonction du temps.

Cashflow evolution : activez cette option pour afficher la courbe présentant l'évolution du cashflow en fonction du temps. Si l'option **segments** est activée, ces valeurs sont affichées pour chaque segment.

Résultats

Statistiques descriptives : les tableaux de statistiques descriptives présentent des statistiques simples pour la variable correspondant aux revenus (RMC). Le nombre d'observations, les valeurs minimales et maximales, les quartiles, la moyenne, la variance et l'écart-type (non biaisé) sont affichés.

Les box plots (ou graphiques boîtes et moustaches) associés sont aussi affichés (Voir description).

CLV : la CLV moyenne est affichée. Si l'option **segments** est ativée, cette valeur est affichée pour chaque segment.

Taux d'attrition estimé : dans ce tableau sont affichés le taux d'attrition et de rétention. Si l'option **segments** est activée, ces valeurs sont affichées pour chaque segment.

Durée estimée avant défection : des statistiques concernant la durée avant défection sont affichées afin de visualiser rapidement la dispersion des temps de départ client. Ainsi, le 1er quartile , le 3e quartile, la médiane et la moyenne des temps de départs clients sont affichés. Si l'option **segments** est activée, ces valeurs sont affichées pour chaque segment.

Evolution du cashflow : ce tableau affiche l'évolution du cashflow. Chaque ligne de ce tableau contient un prévisionnel du cashflow par période. La première ligne du tableau correspond à la période suivant la période la plus récente contenue dans les données d'entrées.

Résultats individuels : la CLV de chaque client est affichée.

CLV prévisionnelle : une simulation sur la CLV moyenne des clients restant dans la base après la dernière date d'attrition enregistrée est réalisée sur la période choisie par l'utilisateur. Si l'option **segments** est activée, cette valeur est affichée pour chaque segment.

Analyse de sensibilité : l'impact d'une augmentation du taux de rétention sur la CLV est affiché. Les variations considérées sont des incréments de 5% à partir du taux de rétention estimé/renseigné. Chaque ligne du tableau correspond à un taux de rétention simulé. La CLV et la durée moyenne avant défection sont affichées en colonne. Si l'option **segments** est activée, ces valeurs sont affichées dans un tableau distinct pour chaque segment.

le graphique **CLV par segment** n'est disponible que lorsque l'option **segments** est active. Sur ce graphique sont affichés sous forme de diagramme en bâtons la CLV pour chaque segment.

Durée estimée avant défection : ce graphique résume sous la forme d'un diagramme en bâtons l'information contenue dans le tableau correspondant décrit plus haut.

Probabilités de churn : activez cette option pour afficher la courbe représentant les probabilités de churn en fonction du temps.

Cashflow evolution : activez cette option pour afficher la courbe représentant l'évolution du cashflow en fonction du temps.

Exemple

Un exemple d'utilisation de la *Customer Lifetime Value* est disponible sur le Centre d'aide XLSTAT à l'adresse :

http://www.xlstat.com/demo-clvf.htm

Bibliographie

Phillip E. Pfeifer, Mark E. Haskins and Robert M. Conroy (2005). Customer Lifetime Value, Customer Profitability, and the Treatment of Acquisition Spending. *Journal of Managerial Issues, Vol. 17(1), pp. 11-25*.

Malthouse, Edward C. (2013). Segmentation and Lifetime value Models Using SAS®. *Cary, NC: SAS institute Inc*.

Customer Long-term Value (CLTV)

À partir de l'historique de vos commandes, la customer long-term value (CLTV) va aider votre entreprise, ou votre association, à déterminer combien rapporteront vos clients et à estimer combien de temps vous les conserverez après acquisition. Aussi elle vous permettera de mieux comprendre le cycle de vie de vos clients, d'identifier des périodes à fort risque de churn, et d'avoir une estimation des profits engendrés par vos clients sur une période.

Dans cette section :

Description

Quelle est la distribution des événements (départs) au cours de la vie d'un client ? Ou encore à quel moment la probabilité de churn est-elle la plus élevée ? Ce sont, entre autres, les questions auxquelles ce module pourrait vous aider à répondre.

Customer Long-term Value (CLTV) :

La CLV, appelée valeur vie client, est un indicateur qui peut être défini comme étant la somme des profits générés par une entreprise tout au long de sa relation avec un client.

Le modèle implémenté ici est le modèle dit de rétention généralisé (GRM). Ce modèle, comme le modèle de rétention simpe (voir la section description de la méthode), est adapté à des données de clients en situation contractuelle.

Dans le modèle de rétention simple (SRM), on suppose que le pourcentage de clients retenus à chaque période, c'est à dire le taux de rétention r, est constant dans le temps. Cependant dans de nombreux cas ce dernier n'est pas toujours constant. Par exemple, certaines organisations, telles que les sociétés de téléphonie ou les fournisseurs d'accès internet, offrent quelques mois de gratuité ou quelques mois avec un prix très réduit avant de passer en tarification pleine. Dans ce genre de cas, on constate en général un taux de rétention assez élevé sur la période couvrant la promotion avant que ce dernier ne commence à s'éffondrer.

Le modèle de rétention généralisé (GRM) est donc une extension du SRM. Ici on suppose que le taux de rétention r varie au cours du temps et que le cashflow m généré par période est tributaire de la période d'annulation.

Cependant, nous considérerons toujours que le fait qu'un client annule au cours d'une période t est indépendant du fait qu'il annule au cours d'une autre période.

Contrairement au SRM, dans lequel on peut estimer la CLV à l'infini car le taux de rétention est constant, ici, on estimera la CLV uniquement à partir de ce qui a été observé dans le passé. Si l'historique client s'étend sur une plage de 10 ans alors le modèle sera capable de produire une estimation de la CLV sur 10 ans grâce à votre historique client.

Modèle de rétention généralisé (GRM) pour le calcul de la CLTV :

Pour la mise en place de ce modèle nous utiliserons des méthodes empruntées à l'analyse de survie, plus précisement l'analyse des tables actuarielles de survie (voir la section description de la méthode).

- **Fonction de rétention** : la fonction de survie qui, dans notre cas correspondra à la fonction de rétention, nous donne pour chacun des intervalles de temps t la probabilité de retenir un client au moins $t-1$ périodes.

Soit R_t l'évènement "le client est retenu en période t" puisque les évènements sont indépendants. On a donc :

$$P(R_1 \bigcap R_2) = P(R_1) \times P(R_2) = r_1 \times r_2$$

En étendant ce raisonnement aux premières $t-1$ périodes, on obtient l'expression de la fonction de rétention :

$$S(t) = P(T \geq t) = \prod_{i=1}^{t-1} r_i$$

Notons que la fonction de churn correpond elle à $1 - S(t)$.

- **Densité de probabilité** : la densité de probabilité correspond à la probabilité qu'a l'entreprise de retenir un client pendant $t-1$ périodes et qu'il annule ou résilie son abonnement pendant la période t. On a :

$$f(t) = P(t = T) = S(t)(1 - r_t) = S(t) - S(t+1)$$

- **Taux de hazard** : il correspond à la probabilité qu'un client résilie son abonnememnt pendant la période t sachant qu'il était présent en période $t-1$. Soit π_t le taux de hazard au temps t

$$\pi_t = P(T = t | T \geq t) = \frac{P(T=t)}{S(t)} \simeq 1 - r_t$$

A noter que le taux de hazard n'est pas toujours une probabilité. Dans notre cas, les intervalles de temps étant discrets, il correspond à une probabilité.

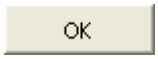
- **Calcul de la CLV** : soit m_t le cashflow (remises éventuelles comprises). Pour un client qui annule durant la période $T = t$, on a $CLV = \sum_{i=0}^{t-1} m_i$. Ici nous chercherons à estimer la valeur attendue de la CLV et donc son espérance. On a donc :

$$E[CLV(T)] = \sum_{t=1}^{\infty} m_{t-1} S(t)$$

La fenêtre de temps T, pour laquelle les valeurs de la fonction de survie $S(t)$ est connue, étant finie, on parlera donc de long-term value. Ce qui explique le terme CLTV pour customer long-term value.

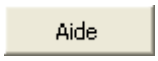
Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondants aux différentes options disponibles, tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.



: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.


: cliquez sur ce bouton pour changer la façon dont vous chargez les données dans XLSTAT. Si le bouton possède une icône de souris, XLSTAT vous permet de faire une sélection à la souris des données. Si le bouton possède une icône de liste, XLSTAT vous permet de sélectionner les données à partir d'une liste. Si le bouton possède une icône orange, des boutons complémentaires avec un point d'interrogation  sont affichés pour vous permettre d'importer les données à partir de fichiers.

Onglet **Général** :

RMC (Revenu Moyen par Client) : sélectionnez les données associées aux revenus. Vous devez sélectionner une seule colonne. Si des en-têtes de colonnes ont été sélectionnées, veuillez vérifier que l'option « Libellés des variables » est activée.

Temps (acquisition/attrition) : sélectionnez dans cet ordre les deux colonnes : date d'acquisition et date d'attrition. Si un client n'est pas encore parti, la valeur corresopondante dans la deuxieme colonne (attrition) doit etre vide. Si des en-têtes de colonnes ont été sélectionnées, veuillez vérifier que l'option « Libellés des variables » est activée.

Segments : activez cette option si vous voulez effectuer des analyses par segments, puis sélectionnez les données qui indiquent à quel segment chaque individu appartient. Si des en-têtes de colonnes ont été sélectionnés, veuillez vérifier que l'option « Libellés des variables » est activée.

Période de souscription : sélectionnez la période de souscription.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des variables : activez cette option si la première ligne des données sélectionnées contient un libellé.

Onglet **Options** :

Libellés des segments : sélectionnez les libellés des segments. Ce champ est uniquement disponible lorsque l'option segments est active et lorsqu'on a au moins une des trois options suivantes d'activée : coûts fixes, taux de remise ou taux de rétention défini par l'utilisateur.

Taux de remise : activez cette option si vous voulez prendre en compte le taux de remise appliqué aux clients. Ce taux est considéré comme fixe. Si l'option segments est activée, sélectionnez une valeur par segment et assurez vous que les lignes soient dans le même ordre que les libellés des segments.

Coûts fixes : activez cette option si vous voulez déduire certains coûts fixes d'exploitation des revenus générés lors du calcul de la CLV. Sélectionnez plusieurs colonnes si vous avez plusieurs coûts à inclure. Si l'option segments est activée, sélectionnez une ligne par segment et assurez-vous que les lignes soient dans le même ordre que les libellés des segments.

Règlement :

- **Début de période** : sélectionnez cette option si le règlement est effectué en début de période.
- **Fin de période** : sélectionnez cette option si le règlement est effectué en fin de période.

Onglet **Sorties** :

Statistiques descriptives : activez cette option pour afficher les statistiques descriptives de la variable qui correspond aux revenus (RMC) et un résumé des actions clients (nombre de clients observés, nombre de clients perdus, nombre de clients pour lesquels nous n'avons pas de date de départ effective (censure)).

Analyse du cycle de vie client : activez cette option pour afficher le tableau d'analyse du cycle de vie des clients. Vous avez le choix entre deux options:

- **Synthèse** : sélectionnez cette option si vous souhaitez afficher un résumé de l'analyse du cycle de vie de vos clients.
- **Complet** : sélectionnez cette option si vous souhaitez avoir des résultats plus complets sur l'analyse du cycle de vie de vos clients.

CLV : activez cette option pour afficher la CLV moyenne. Si l'option **segments** est activée, cette valeur est affichée pour chaque segment.

Evolution de la Customer Long-term Value (CLTV) : activez cette option pour afficher, pour chaque période, la CLV ainsi que la CLTV qui correspond ici à la CLV cumulée jusqu'à la période considérée.

Comparaison des segments : si l'option **segments** est activée, activez cette option pour réaliser une comparaison des fonctions de rétention cumulée des différents segments.

Onglet **Graphiques** :

CLV par segment : si l'option **segments** est activée, activez cette option pour afficher sous forme de diagramme le graphique de la CLV par segments.

Fonction de rétention : activez cette option pour afficher le graphique représentant l'évolution de la fonction de rétention cumulée au cours du temps.

Densité de probabilité : activez cette option pour afficher la courbe représentant la fonction de densité estimée pour chaque période.

Taux de hazard : activez cette option pour afficher la courbe représentant la fonction de hazard estimée pour chaque période.

Evolution de la Customer Long-term Value (CLTV) : activez cette option pour afficher le graphique associé à l'évolution de la CLTV au cours du temps.

Graphiques de comparaison : si les options **Segments** et **Comparaison des segments** sont activées, activez cette option pour afficher les graphiques comparant les courbes de rétention, de densité et de hasard pour les différents segments, si elles ont été selectionnées.

Notons que les périodes pour lesquels des données censurées ont été observées sont identifiées sur le graphique par un "+".

Résultats

Statistiques descriptives : les tableaux de statistiques descriptives présentent des statistiques simples pour la variable correspondant aux revenus (RMC). Le nombre d'observations, les valeurs minimales et maximales, les quartiles, la moyenne, la variance et l'écart-type (non biaisé) sont affichés. Aussi un résumé des actions clients (nombre de clients observés, nombre de clients perdus, nombre de clients pour lesquels nous n'avons pas de date de départ effective (censure)) est affiché.

Les box plots (ou graphiques boîtes à moustaches) associés sont aussi affichés.

CLV : la CLV moyenne est affichée. Si l'option **segments** est ativée, cette valeur est affichée pour chaque segment.

Analyse du cycle de vie client : dans ce tableau sont affichés les résultats suivants :

- **Synthèse** :
- **Période** : intervalle de temps.
- **Nb. clients** : nombre de clients présents pendant l'intervalle de temps.
- **Nb. perdus** : nombre de client perdus pendant l'intervalle de temps.
- **Censurées** : nombre de clients présents à l'issue de la période lorsque celle-ci correspond à la date de fin de l'étude.
- **Effectivement à risque** : nombre de clients considérés comme étant à risque pendant l'intervalle de temps.
- **Taux de rétention** : proportion de clients présents pendant l'intervalle de temps.
- **Taux d'attrition** : proportion de clients perdus pendant l'intervalle de temps.
- **Complet** : en plus de ceux présents dans le résumé nous avons les indicateurs suivants :
- **Taux de rétention cumulé** : probabilité qu'a un client de rester client au moins jusqu'au temps considéré.
- **Taux de churn cumulé** : cumul du taux d'attrition jusqu'au temps considéré.
- **Densité de probabilité** : fonction de densité estimée au milieu de l'intervalle de temps considéré.
- **Taux de hasard** : estimation du taux de hasard au milieu de l'intervalle de temps considéré.

Temps de rétention médian : vous trouverez dans ce tableau le temps médian résiduel de rétention au début de l'expérience, ainsi que l'écart-type de ce dernier. Cette statistique permet d'évaluer le temps au bout duquel le nombre de clients étudiés a réduit de moitié.

Si l'option **segments** est activée, ces valeurs sont affichées pour chaque segment.

Comparaison des segments : vous trouverez dans ce tableau une comparaison des fonctions de rétention cumulée des différents segments.

Ce tableau affiche les statistiques correspondants à trois tests : le Log-rank test, le test de Wilcoxon, et le test de Tarone Ware test. Ces tests s'appuient tous sur le test du Khi². Plus la p-value est faible, plus la différence entre les courbes est significative.

Si la p-value obtenue par le test du log-rank est significative au seuil $alpha = 5\%$, des tests de comparaisons multiples sont effectués sur les segments 2 à 2. Nous utilisons, dans ce cas, le test de Dunn-Sidak qui est un dérivé du test de Bonferroni et qui s'avère plus performant dans certaines situations.

$$\alpha' = 1 - (1 - \alpha)^{1/g}.$$

où g est le nombre de segments comparés.

Le graphique **CLV par segment** n'est disponible que lorsque l'option **segments** est active. Sur ce graphique sont affichés, sous forme de diagramme en bâtons, la CLV pour chaque segment.

Exemple

Un exemple d'utilisation de la *Customer long-term Value* est disponible sur le Centre d'aide XLSTAT à l'adresse :

http://www.xlstat.com/demo-cltvf.htm

Bibliographie

Phillip E. Pfeifer, Mark E. Haskins and Robert M. Conroy (2005). Customer Lifetime Value, Customer Profitability, and the Treatment of Acquisition Spending. *Journal of Managerial Issues*, Vol. 17(1), pp. 11-25.

Malthouse, Edward C. (2013). Segmentation and Lifetime value Models Using SAS®. *Cary, NC: SAS institute Inc*.

Process : modération et médiation

Utilisez cet outil de manière complémentaire à la régression linéaire pour approfondir la compréhension du phénomène étudié et ainsi répondre aux questions "quand et comment" cet effet se produit.

Dans cette section :

Description

La méthode Process a été popularisée par Andrew F. Hayes en 2013. Celle-ci est complémentaire de l'approche PLS-SEM et qui est très populaire dans le domaine du marketing ainsi que dans celui des sciences sociales et comportementales. Process se décompose autour de deux notions principales : la médiation et la modération.

L'analyse de médiation est utilisée pour tester des hypothèses éclairant les divers mécanismes intermédiaires par lesquels les effets causaux surviennent, alors que l'analyse de modération est utilisée pour explorer les questions relatives aux conditions d'un effet.

Le point commun de ces deux notions est de vouloir explorer le rôle joué par une troisième variable dans la relation entre une variable explicative X et une variable réponse Y.

Modèle de médiation

Le modèle de médiation suppose que X influence un médiateur M qui, à son tour, influence Y . Si X à un effet sur Y via M, alors le système suivant résume les relations entre X, M et Y : $$ \left\lbrace \begin{aligned} & Y = i_Y + cX +bM +\epsilon_Y \ & M = i_M + aX +\epsilon_M \ \end{aligned} \right. $$ Les différents paramètres de ces équations s'estiment par la méthode des moindres carrés et permettent d'obtenir l'effet indirect (qui représente la manière dont Y est influencée par X à travers M) et l'effet direct (symbolisant le chemin qui mène de X à Y sans passer par le médiateur M) propres au modèle : $$ Effet_{Direct}=c $$ $$ Effet_{Indirect}=ab $$ Les effets directs et indirects permettent de conclure sur la significativité ou non du modèle de médiation.

Modèle de modération

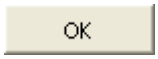
Le modèle de modération suppose que X influence Y plus ou moins fortement en fonction d'un modérateur W. Si X à un effet sur Y modulé par W, alors l'équation suivante résume les relations entre X, W et Y : $$ Y = i + dX +eW + fInter+\epsilon_Y $$ avec $Inter$ qui représente la variable d'intéraction composée à partir des variables X et W.

Les différents paramètres de cette équation s'estiment par la méthode des moindres carrés.

La significativité de la modération repose sur la significativité ou non du coefficient f associé à la variable d'intéraction.

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.


: cliquez sur ce bouton pour changer la façon dont vous chargez les données dans XLSTAT. Si le bouton possède une icône de souris, XLSTAT vous permet de faire une sélection à la souris des données. Si le bouton possède une icône de liste, XLSTAT vous permet de sélectionner les données à partir d'une liste. Si le bouton possède une icône orange, des boutons complémentaires avec un point d'interrogation  sont affichés pour vous permettre d'importer les données à partir de fichiers.

Y / Variable dépendante : sélectionnez la variable réponse quantitative que vous souhaitez modéliser. Si des en-têtes de colonnes ont été sélectionnés, veuillez vérifier que l'option « Libellés des variables » est activée.

X / Variable explicative : sélectionnez la variable explicative quantitative sur la feuille Excel. Les données sélectionnées doivent être de type numérique. Si le libellé des variables a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

M / Variable(s) médiatrice(s) : sélectionnez la ou les variable(s) médiatrice(s) quantitative(s) sur la feuille Excel. Les données sélectionnées doivent être de type numérique. Si le libellé des variables a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

W / Variable modératrice : sélectionnez la variable modératrice quantitative sur la feuille Excel. Les données sélectionnées doivent être de type numérique. Si le libellé des variables a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

V / Variable modératrice : sélectionnez la variable modératrice quantitative sur la feuille Excel. Les données sélectionnées doivent être de type numérique. Si le libellé des variables a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

Covariables : activez cette option si vous voulez inclure une ou plusieurs variables explicatives au modèle.

Libellés des variables : activez cette option si la première ligne des données sélectionnées contient un libellé.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Numéro du modèle : choisissez le modèle à utiliser pour les calculs (le modèle est représenté sur la partie droite de la boîte de dialogue).

Intervalle de confiance (%) : entrez l'étendue en pourcentage de l'intervalle de confiance à utiliser pour le calcul des intervalles de confiance autour des paramètres. Valeur par défaut : 95.

Rééchantillonnages : entrez le nombre d'échantillons à générer lors du bootstrap.

Statistiques descriptives : activez cette option pour afficher les statistiques descriptives pour les variables sélectionnées.

Matrice de corrélation : activez cette option pour afficher un aperçu des corrélations entre les différentes variables sélectionnées.

Graphique de Johnson-Neyman : activez cette option pour afficher le graphique de Johnson-Neyman.

Graphique de l'effet conditionnel : activez cette option pour afficher le graphique montrant l'évolution de l'effet conditionnel dans le modèle.

Résultats

Statistiques descriptives : le tableau des statistiques descriptives présente des statistiques simples pour toutes les variables sélectionnées. Le nombre de valeurs manquantes, le nombre de valeurs non manquantes, la moyenne, l'écart-type sont affichés pour les variables quantitatives.

Matrice de corrélation : ce tableau est affiché afin de vous permettre d'avoir un aperçu des corrélations entre les différentes variables sélectionnées.

Coefficients d'ajustement : dans ce tableau sont affichées les statistiques relatives à l'ajustement du modèle de régression :

- **Observations** : le nombre d'observations prises en compte dans les calculs. Dans les formules présentées ci-dessous n désigne le nombre d'observations.
- **Somme des poids** : la somme des poids des observations prises en compte dans les calculs. Dans les formules présentées ci-dessous W désigne la somme des poids.
- **DDL** : le nombre de degrés de liberté pour le modèle retenu (correspondant à la partie erreurs).
- **R²** : le coefficient de détermination du modèle. La valeur de ce coefficient est comprise entre 0 et 1. XLSTAT le calcule comme suit :

$$R^2 = \frac{\sum_{i=1}^{n} w_i(\hat{y}_i - \bar{y}_i)^2}{\sum_{i=1}^{n} w_i(y_i - \bar{y}_i)^2} \text{ avec } \bar{y} = \frac{1}{W}\sum_{i=1}^{n} w_i y_i$$

Le R^2 s'interprète comme la proportion de la variabilité de la variable dépendante expliquée par le modèle. Plus le R^2 est proche de 1, meilleur est le modèle. L'inconvénient du R^2 est qu'il ne prend pas en compte le nombre de variables utilisées pour ajuster le modèle.

- **R² ajusté** : le coefficient de détermination ajusté du modèle. Le R^2 ajusté peut être négatif si le R^2 est voisin de zéro. Ce coefficient n'est affiché que si la constante du modèle n'est pas fixée par l'utilisateur. Sa valeur est définie par

$$\hat{R}^2 = 1 - (1 - R^2)\frac{W - 1}{W - p - 1}$$

Le R^2 ajusté est une correction du R^2 qui permet de prendre en compte le nombre de variables utilisées dans le modèle.

- **MCE** : la moyenne des carrés des erreurs (MCE) est définie par :

$$MCE = \frac{1}{W - p^*}\sum_{i=1}^{n} w_i(y_i - \hat{y}_i)^2$$

- **RMCE** : la racine de la moyenne des carrés des erreurs (RMCE) est la racine carrée de la MCE.
- **MAPE** : la *Mean Absolute Percentage Error* est calculée comme suit :

$$MAPE = \frac{100}{W}\sum_{i=1}^{n} w_i \left|\frac{y_i - \hat{y}_i}{y_i}\right|$$

- **DW** : le coefficient de Durbin-Watson est défini par

$$DW = \frac{\sum_{i=2}^{n}[(y_i - \hat{y}_i) - (y_{i-1} - \hat{y}_{i-1})]^2}{\sum_{i=1}^{n} w_i(y_i - \hat{y}_i)^2}$$

Ce coefficient correspond au coefficient d'autocorrélation d'ordre 1 et permet de vérifier que les résidus du modèle ne sont pas autocorrélés, sachant que l'indépendance des résidus est l'une des hypothèses de base de la régression linéaire. L'utilisateur pourra se référer à une table des coefficients de Durbin-Watson pour vérifier si l'hypothèse d'indépendance des résidus est acceptable.

- **Cp** : le coefficient Cp de Mallows est défini par

$$Cp = \frac{SCE}{\hat{\sigma}} + 2p^* - W$$

où SCE est la somme du carré des erreurs pour le modèle avec p variables explicatives, et où $\hat{\sigma}$ correspond à l'estimateur de la variance des résidus pour le modèle comprenant toutes les variables explicatives. Plus le coefficient Cp est proche de p* moins le modèle est biaisé.

- **AIC** : le critère d'information d'Akaike (Akaike's Information Criterion) est défini par

$$AIC = W \ln(\frac{SCE}{W}) + 2p^*$$

Ce critère proposé par Akaike (1973) dérive de la théorie de l'information, et s'appuie sur la mesure de Kullback et Leibler (1951). C'est un critère de sélection de modèles qui pénalise les modèles pour lesquels l'ajout de nouvelles variables explicatives n'apporte pas suffisamment d'information au modèle, l'information étant mesurée au travers de la SCE. On cherche à minimiser le critère AIC.

- **SBC** : le critère bayésien de Schwarz (Schwarz's Bayesian Criterion) est défini par

$$SBC = W \ln(\frac{SCE}{W}) + \ln(W)p^*$$

Ce critère proposé par Schwarz (1978) est proche du critère AIC, et comme ce dernier on cherche à le minimiser.

- **PC** : le critère de prédiction d'Amemiya (Amemiya's Prediction Criterion) est défini par

$$PC = \frac{(1 - R^2)(W + p^*)}{W - p^*}$$

Le **tableau d'analyse de la variance** permet d'évaluer le pouvoir explicatif des variables explicatives. Le pouvoir explicatif est évalué en comparant l'ajustement (au sens des moindres carrés) du modèle final avec l'ajustement du modèle rudimentaire composé d'une constante égale à la moyenne de la variable dépendante.

Le tableau des **paramètres du modèle** affiche l'estimation des paramètres, l'écart-type correspondant, le t de Student, la probabilité correspondante, ainsi que l'intervalle de confiance. Si l'intervalle de confiance comprend 0, alors le poids d'une variable dans le modèle n'est pas significatif.

Le **tableau effet direct de X sur Y** affiche l'estimation de l'effet direct de X sur Y, l'écart-type correspondant, le t de Student, la probabilité correspondante, ainsi que l'intervalle de confiance.

Le **tableau effet direct conditionnel de X sur Y** affiche l'estimation de l'effet direct conditionnel de X sur Y pour trois valeurs du modérateur (le 16ème centile, la médiane, et le 84ème centile), l'écart-type correspondant, le t de Student, la probabilité correspondante, ainsi que l'intervalle de confiance.

Le **tableau effet indirect de X sur Y** affiche l'estimation de l'effet indirect de X sur Y, ainsi que l'intervalle de confiance et l'écart-type correspondant qui sont obtenus par la méthode du bootstrap. Si l'intervalle de confiance comprend 0, alors l'effet indirect de X sur Y dans le modèle n'est pas significatif.

Le **tableau effet indirect conditionnel de X sur Y** affiche l'estimation de l'effet indirect conditionnel de X sur Y pour trois valeurs du modérateur (le 16ème centile, la médiane, et le 84ème centile), ainsi que l'intervalle de confiance et l'écart-type correspondant qui sont obtenus par la méthode du bootstrap. Si l'intervalle de confiance comprend 0, alors, pour la valeur du modérateur en question, l'effet indirect conditionnel de X sur Y dans le modèle n'est pas significatif.

Le **tableau indice de médiation modérée** affiche l'estimation de l'indice de médiation modérée, ainsi que l'intervalle de confiance et l'écart-type correspondant qui sont obtenus par la méthode du bootstrap. Si l'intervalle de confiance comprend 0, alors le modèle de médiation modérée n'est pas considéré comme étant significatif.

Graphique de Johnson-Neyman : ce graphique permet de visualiser à partir de quelle valeur du modérateur l'effet devient significatif.

Graphique de l'effet conditionnel : ce graphique permet de visualiser l'évolution de l'effet conditionnel dans le modèle pour trois valeurs du modérateur (le 16ème centile, la médiane, et le 84ème centile).

Exemple

Un tutoriel sur la façon d'utiliser la fonctionnalité "Process : modération et médiation" est disponible sur le Help Center de XLSTAT:

http://www.xlstat.com/demo-processf.htm

Bibliographie

Aiken, L. S., & West, S. G. (1991). Multiple regression: Testing and interpreting interactions. Sage Publications: Thousand Oaks, CA.

Hayes, A. F. (2018). Introduction to Mediation, Moderation, and Conditional Process Analysis: A Regression-Based Approach (2 ed.). Guilford Press: New York, NY.

Preacher, K. J., Rucker, D. D., & Hayes, A. F. (2007). Addressing moderated mediation hypotheses: Theory, methods, and prescriptions. Multivariate Behavioral Research, 42(1), 185–227.

Analyse conjointe

Plans d'expériences pour l'analyse conjointe

Utilisez cet outil pour générer les plans d'expériences associés à une analyse conjointe par profils complets.

Dans cette section :

Description

Le principe de l'analyse conjointe est de présenter un ensemble de produits (appelés aussi profils) à des individus qui devront soit les noter, soit les classer, soit en choisir certains.

Dans un cadre « idéal », il faudrait que les individus testent tous les produits possibles. Or, cela est très vite impossible, les capacités de chacun étant limitées et le nombre de combinaisons augmentant très rapidement avec le nombre d'attributs (si on veut étudier 5 attributs avec 3 modalités chacun, on a déjà 243 produits possibles). On utilise donc des méthodes de plans d'expériences afin d'obtenir un nombre de profils acceptable pour les juges tout en gardant de bonnes propriétés statistiques.

XLSTAT permet de générer plusieurs plans uniques, ce qui est un avantage notamment lorsque l'on souhaite interroger un grand nombre de personnes. Le nombre de combinaisons différentes étant plus important, les plans de l'analyse seront plus robustes dans l'analyse des effets. De plus, le fait d'inclure des plans différents réduit l'impact du contexte psychologique et des effets d'ordre.

XLSTAT-Conjoint propose deux méthodes d'analyse conjointe différentes : les profils complets et l'analyse basée sur le choix.

Les profils complets

La première étape de l'analyse conjointe nécessite le choix d'un certain nombre de facteurs décrivant un produit. Ces facteurs doivent être qualitatifs. Ainsi, par exemple, si on cherche à introduire un nouveau produit sur un marché, on pourra choisir comme facteurs différenciateurs : son prix, sa qualité, sa longévité… et pour chaque facteur, il faudra définir un

certain nombre de modalités (différents prix, différentes durées de vie…). Cette première étape est primordiale et se fera à l'aide des experts du marché étudié.

Une fois cette première étape passée, il faut essayer de comprendre le mécanisme de choix d'un produit plutôt qu'un autre. Pour cela on va proposer un certain nombre de produits (combinant des modalités différentes des facteurs étudiés). On ne pourra pas proposer tous les produits possibles, on sélectionnera donc des produits en utilisant des plans d'expériences avant de les présenter à des individus qui devront les noter ou les classer.

Il s'agit alors de la méthode des profils complets. Elle est la plus ancienne des méthodes d'analyse conjointe, on cherche à construire un plan d'expérience comprenant un nombre limité de profils complets que chaque individu interrogé devra ensuite classer ou noter.

XLSTAT-Conjoint utilise les plans factoriels fractionnaires afin de générer les profils qui seront ensuite présentés aux personnes interrogées. Quand aucun plan n'est disponible ou si vous préférez, XLSTAT-Conjoint peut utiliser des algorithmes de recherche de plans D-optimaux (voir description des plans d'effet de facteurs).

Dans le cadre de l'analyse conjointe classique, les questionnaires utilisés sont basés sur la notation ou le classement d'un certain nombre de profils complets.

Il faut sélectionner des attributs d'intérêt pour le produit et les modalités associées à ces attributs. XLSTAT-Conjoint génère ensuite les profils à classer / noter pour chaque répondant.

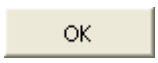
Les combinaisons interdites

Pour certaines conceptions de plans d'expérience, des combinaisons de facteurs ne sont pas réalisables. Cela peut être dû à différentes raisons : équipements, produits,… Dans ces cas, il est possible d'indiquer ces combinaisons interdites, le plan d'expérience généré ne les prendra alors pas en compte.

En choisissant d'ajouter des combinaisons interdites, le plan généré sera forcément un plan optimisé.

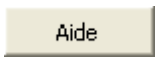
Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.



: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.

 : cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.

 : cliquez sur ce bouton pour changer la façon dont vous chargez les données dans XLSTAT. Si le bouton possède une icône de souris, XLSTAT vous permet de faire une sélection à la souris des données. Si le bouton possède une icône de liste, XLSTAT vous permet de sélectionner les données à partir d'une liste.

Onglet **Général**:

Tableau facteurs/modalités : entrez le tableau regroupant le nom des facteurs et leurs modalités.

Nombre maximal de profils : entrez le nombre maximal de profils qui seront affichés et qui devront être classés par les répondants.

Nombre de réponses : entrez le nombre de répondants dans votre analyse.

Nombre de cas d'exclusion : entrez le nombre de cas d'exclusion. Les cas d'exclusion sont des cas évalués par les répondants, mais qui ne sont pas incluses dans l'analyse conjointe ensuite. Ces cas sont intéressant à ajouter à votre modèle car ils vont permettre de vérifier sa validité. Ils sont générés de façon aléatoire.

- **Mélanger aléatoirement avec les autres cas** : mélange de façon aléatoire les cas d'exclusion avec les autres cas du plan d'expérience.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Onglet **Options**:

Nombre de plans : Cochez cette option si vous souhaitez générer plusieurs plans.

- Répondants par plan : Entrez un vecteur de la taille du nombre de plan, contenant le nombre de répondants par plan souhaité.

Combinaisons interdites : cochez cette option si vous souhaitez rentrer des combinaisons interdites.

Plan d'expériences :

- Plan D-Optimal : sélectionnez cette option pour générer un plan D-Optimal. Ce plan correspondra exactement aux facteurs sélectionnés (voir le chapitre sur les plans d'effet de facteurs dans l'aide pour plus de précisions).
- Plan orthogonal : sélectionnez cette option afin de trouver un plan orthogonal proche des paramètres entrés par l'utilisateur et présent dans la base de données XLSTAT.

Conditions d'arrêt : le nombre d'itérations et l'indice de convergence pour l'obtention d'un plan d'expérience peuvent être modifiés.

Onglet **Sorties** :

Bilan de l'optimisation : activez cette option pour afficher le bilan de l'optimisation effectuée pour générer le plan d'expérience.

Tableau de Burt : activez cette option pour afficher le tableau de Burt associé au plan d'expériences.

Plan codé : activez cette option pour afficher le tableau du plan d'expériences encodé dans le cas d'un plan d-optimal.

Afficher les feuilles individuelles : activez cette option si vous voulez qu'une feuille pour chaque répondant soit générée afin de l'utiliser à titre de feuille de réponse.

Affectation : si l'option précédente est activée, choisissez si vous voulez que les profils soient présentés toujours dans le même ordre (fixe) ou dans des ordres aléatoires pour les individus interrogés (aléatoire).

Inclure des références : si l'option d'affichage des feuilles individuelles est activée, activez cette option si vous voulez que les feuilles du classeur soient liées en utilisant des formules Excel. Lorsqu'on entre une réponse dans la feuille individuelle, elle est automatiquement ajoutée au tableau général.

Boite de dialogue Combinaisons interdites :

Cette boite de dialogue vous permet de sélectionner des combinaisons interdites. Pour cela, sélectionnez dans la partie de gauche les combinaisons de modalités interdites, puis cliquez sur le bouton ajouter. Les combinaisons vont alors s'afficher dans la partie de droite. Il est possible de supprimer une ou plusieurs combinaisons sélectionnées en cliquant sur celles-ci puis sur le bouton supprimer. Une fois que vous avez sélectionné les combinaisons interdites, cliquez sur le bouton OK.

Boite de dialogue Plans pour l'analyse conjointe :

Sélection du plan d'expérience : Cette boîte de dialogue vous permet de sélectionner le plan d'expérience que vous désirez utiliser. Ainsi, une liste de plans factoriels fractionnaires est présentée avec leur distance respective au plan qui devait être généré. Si vous sélectionnez un plan et que vous cliquez sur sélectionner, c'est le plan sélectionné qui apparaitra.

Résultats

Informations sur les variables : dans ce tableau sont récapitulées toutes les informations sur les facteurs sélectionnés.

Informations sur les combinaisons interdites : dans ce tableau sont récapitulées l'ensemble des combinaisons interdites sélectionnées.

Plan d'analyse conjointe : dans ce tableau sont affichés les profils générés. Les colonnes vide correspondent aux réponses à remplir par les individus interrogés suite au questionnaire (classement / ordre). Si l'option « afficher les feuilles individuelles » est activée et que l'option « inclure des références » l'est aussi. Alors les colonnes vides du tableau font directement références aux feuilles associées à chaque individu.

Lancer l'analyse : Une fois que tous les individus ont rempli le plan d'analyse conjointe, vous pouvez cliquer sur le bouton « Lancer l'analyse » afin d'ouvrir la boite de dialogue préremplie permettant d'effectuer l'analyse conjointe.

Plan codé : le plan encodé est affiché. Ce tableau n'est disponible que si le plan d'expériences est un plan d-optimal.

Tableau de Burt : le tableau de Burt est affiché si l'option correspondante a été activée dans la boîte de dialogue. Une visualisation 3D de ce tableau est aussi affichée si l'option est activée dans l'onglet « Graphiques » de la boîte de dialogue.

Bilan de l'optimisation : dans ces tableaux sont affichés les détails de l'optimisation pour la construction du plan d'expérience. Ce bilan présente le nombre d'itérations nécessaires à la réalisation du plan ainsi que la D-efficacité ou la diagonalité. Dans le cas de plans simples, la D-éfficatité est affichée, c'est un indicateur relatif d'efficacité, permettant la comparaison avec d'autres plans de même taille, le but est de maximiser cet indicateur qui est entre 0 et 1. La diagonalité est affichée lorsque l'on souhaite construire des plans multiples, c'est une mesure qui permet de mesurer que la confusion entre facteurs et interactions est minimale. Cette mesure est comprise entre 0 et 1, plus elle est proche de 1, plus la confusion est faible, ce qui est recherché. Le plan sélectionné au final est mis en gras dans le tableau.

Feuilles Nom du modèle_Res : ces feuilles sont les feuilles qui s'affichent lorsque l'option « afficher les feuilles individuelles » est activée. Chaque feuille correspond à un questionnaire avec le nom de l'analyse, le numéro du répondant ainsi qu'un tableau à remplir par le répondant (la dernière colonne doit être remplie).

Exemple

Un exemple d'analyse conjointe basée sur les profils complets est disponible sur le Centre d'aide XLSTAT :

http://www.xlstat.com/demo-conjointf.htm

Bibliographie

Green, P.E. and Srinivasan, V. (1990). Conjoint analysis in Marketing: New Developments with implication for research and practice, Journal of Marketing, **54** (4), 3-19.

Gustafson, A., Herrmann, A. and Huber F. (eds.) (2001). Conjoint Measurement. Method and Applications, Springer.

Plans d'expériences pour l'analyse conjointe basée sur le choix

Utilisez cet outil pour générer des plans d'expériences associés à une analyse conjointe basée sur le choix (CBC).

Dans cette section :

Description

Le principe de l'analyse conjointe basée sur le choix est de présenter des groupes de produits et de demander aux individus interrogés de choisir entre les différents produits proposés. Ces groupes sont générés en utilisant des plans d'expériences. Ce processus de création se sépare en deux phases :

- le nombre de produits générés dans un premier temps, l'est en utilisant la même méthode que pour l'analyse en profils complets. Des plans factoriels fractionnaires ou optimisés sont utilisés pour générer des profils.
- Une fois ces profils générés, ils sont répartis dans des groupes de choix parmi lesquels le répondant va devoir choisir (ou alors décider de ne pas choisir). Des plans en blocs incomplets équilibrés sont utilisés pour cette étape (voir l'aide sur les plans d'expériences pour l'analyse sensorielle du module XLSTAT-MX).

XLSTAT permet de générer plusieurs plans uniques, ce qui est un avantage notamment lorsque l'on souhaite interroger un grand nombre de personnes. Le nombre de combinaisons différentes étant plus important, les plans de l'analyse seront plus robustes dans l'analyse des effets. De plus, le fait d'inclure des plans différents réduit l'impact du contexte psychologique et des effets d'ordre.

L'analyse conjointe basée sur le choix (CBC)

La première étape de l'analyse conjointe nécessite le choix d'un certain nombre de facteurs décrivant un produit. Ces facteurs doivent être qualitatifs. Ainsi, par exemple, si on cherche à introduire un nouveau produit sur un marché, on pourra choisir comme facteurs différenciateurs : son prix, sa qualité, sa longévité… et pour chaque facteur, il faudra définir un

certain nombre de modalités (différents prix, différentes durées de vie…). Cette première étape est primordiale et se fera à l'aide des experts du marché étudié.

Une fois cette première étape passée, il faut comprendre le mécanisme de choix d'un produit plutôt qu'un autre. Pour cela on va proposer un certain nombre de produits (combinant des modalités différentes des facteurs étudiés). On ne pourra pas proposer tous les produits possibles, on sélectionnera donc des produits en utilisant des plans d'expériences.

La théorie de l'analyse conjointe a montré que l'analyse conjointe basée sur les profils complets était moins efficace que l'analyse conjointe basée sur le choix qui, au lieu de présenter des profils à classer ou à noter, va demander aux individus de choisir un profil plutôt qu'un autre ou même aucun profil.

Cette génération donnera donc deux plans d'expériences. L'un rassemblant les profils et le second présentant ces mêmes profils par groupes (le nombre d'élément dans chaque groupe est à paramétrer par l'utilisateur).

XLSTAT-Conjoint utilise les plans factoriels fractionnaires afin de générer les profils et ensuite des plans en blocs incomplets équilibrés pour générer les choix.

XLSTAT-Conjoint permet aussi d'ajouter l'option zéro, c'est-à-dire l'option de non choix.

XLSTAT-Conjoint permet d'obtenir un tableau global mais aussi des tableaux pour chaque individu séparément. Il suffira à celui-ci de remplir une case dans un tableau Excel afin que sa réponse soit comptabilisée dans l'analyse.

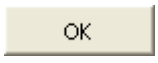
Les combinaisons interdites

Pour certaines conceptions de plans d'expérience, des combinaisons de facteurs ne sont pas réalisables. Cela peut être dû à différentes raisons : équipements, produits,… Dans ces cas, il est possible d'indiquer ces combinaisons interdites, le plan d'expérience généré ne les prendra alors pas en compte.

En choisissant d'ajouter des combinaisons interdites, le plan généré sera forcément un plan optimisé.

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.

 : cliquez sur ce bouton pour effacer les sélections de données.

 : cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.

 : cliquez sur ce bouton pour changer la façon dont vous chargez les données dans XLSTAT. Si le bouton possède une icône de souris, XLSTAT vous permet de faire une sélection à la souris des données. Si le bouton possède une icône de liste, XLSTAT vous permet de sélectionner les données à partir d'une liste.

Onglet **Général** :

Tableau facteurs/modalités : entrez le tableau regroupant le nom des facteurs et leurs modalités.

Nombre maximal de profils : entrez le nombre maximal de profils qui seront affichés et qui devront être classés par les répondants.

Nombre de réponses : entrez le nombre de répondants dans votre analyse.

Nombre maximum de comparaisons : entrez le choix à effectuer par chaque répondant (ce nombre doit être plus grand que le nombre de profils).

Nombre de profils par comparaison : entrez le nombre de profils présenté lors de chaque choix.

Nombre de cas d'exclusion : entrez le nombre de cas d'exclusion. Les cas d'exclusion sont des cas évalués par les répondants, mais qui ne sont pas incluses dans l'analyse conjointe ensuite. Ces cas sont intéressant à ajouter à votre modèle car ils vont permettre de vérifier sa validité. Ils sont générés de façon aléatoire.

- **Mélanger aléatoirement avec les autres cas** : mélange de façon aléatoire les cas d'exclusion avec les autres cas du plan d'expérience.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Onglet **Options** :

Nombre de plans : Cochez cette option si vous souhaitez générer plusieurs plans.

- Répondants par plan : Entrez un vecteur de la taille du nombre de plan, contenant le nombre de répondants par plan souhaité.
- Comparaisons par plan : Entrez un vecteur de la taille du nombre de plans, contenant le nombre de combinaisons par plan souhaité.

Combinaisons interdites : cochez cette option si vous souhaitez rentrer des combinaisons interdites.

Plan d'expériences :

- Plan D-Optimal : sélectionnez cette option pour générer un plan D-Optimal. Ce plan correspondra exactement aux facteurs sélectionnés (voir le chapitre sur les plans d'effet de facteurs dans l'aide pour plus de précisions).
- Plan orthogonal : sélectionnez cette option afin de trouver un plan orthogonal proche des paramètres entrés par l'utilisateur et présent dans la base de données XLSTAT.

Conditions d'arrêt : le nombre d'itérations et l'indice de convergence pour l'obtention d'un plan d'expérience peuvent être modifiés.

Onglet **Sorties** :

Bilan de l'optimisation : activez cette option pour afficher le bilan de l'optimisation effectuée pour générer le plan d'expérience.

Tableau de Burt : activez cette option pour afficher le tableau de Burt associé au plan d'expériences.

Plan codé : activez cette option pour afficher le tableau du plan d'expériences encodé dans le cas d'un plan d-optimal.

Afficher les feuilles individuelles : activez cette option si vous voulez qu'une feuille pour chaque individu interrogé soit générée afin de l'utiliser à titre de feuille de réponse.

Affectation : si l'option précédente est activée, choisissez si vous voulez que les choix soient présentés toujours dans le même ordre (fixe) ou dans des ordres aléatoires pour les répondants (aléatoire).

Inclure des références : si l'option d'affichage des feuilles individuelles est activée, activez cette option si vous voulez que les feuilles du classeur soient liées en utilisant des formules Excel. Lorsqu'on entre une réponse dans la feuille individuelle, elle est automatiquement ajoutée au tableau général.

Inclure l'option zéro : si l'option « Afficher les feuilles individuelles » est activée, activez cette option si vous voulez inclure une option de non choix dans les feuilles individuelles générées.

Boite de dialogue **Combinaisons interdites** :

Cette boite de dialogue vous permet de sélectionner des combinaisons interdites. Pour cela, sélectionnez dans la partie de gauche les combinaisons de modalités interdites, puis cliquez sur le bouton ajouter. Les combinaisons vont alors s'afficher dans la partie de droite. Il est possible

de supprimer une ou plusieurs combinaisons sélectionnées en cliquant sur celles-ci puis sur le bouton supprimer. Une fois que vous avez sélectionné les combinaisons interdites, cliquez sur le bouton OK.

Boite de dialogue **Plans pour l'analyse conjointe** :

Sélection du plan d'expérience : Cette boîte de dialogue vous permet de sélectionner le plan d'expérience que vous désirez utiliser. Ainsi, une liste de plans factoriels fractionnaires est présentée avec leur distance respective au plan qui devait être généré. Si vous sélectionnez un plan et que vous cliquez sur sélectionner, c'est le plan sélectionné qui apparaitra.

Résultats

Informations sur les variables : dans ce tableau sont récapitulées toutes les informations sur les facteurs sélectionnés.

Informations sur les combinaisons interdites : dans ce tableau sont récapitulées l'ensemble des combinaisons interdites sélectionnées.

Profils : dans ce tableau sont affichées les profils générés.

Plan d'analyse conjointe : dans ce tableau sont affichés les choix générés. Pour chaque choix, le numéro du profil associé est donné. Les colonnes vides servent à être remplies avec les réponses. L'individu interrogé entrera soit le choix sélectionné (numéro associé à la colonne choisie), soit zéro, si l'individu a sélectionné le non choix. Si l'option « afficher les feuilles individuelles » est activée et que l'option « inclure des références » l'est aussi. Alors les colonnes vides du tableau font directement références aux feuilles associées à chaque individu.

Lancer l'analyse : Une fois que tous les individus ont rempli le plan d'analyse conjointe, vous pouvez cliquer sur le bouton « Lancer l'analyse » afin d'ouvrir la boite de dialogue préremplie permettant d'effectuer l'analyse conjointe.

Plan codé : le plan encodé est affiché. Ce tableau n'est disponible que si le plan d'expériences est un plan d-optimal.

Tableau de Burt : le tableau de Burt est affiché si l'option correspondante a été activée dans la boîte de dialogue. Une visualisation 3D de ce tableau est aussi affichée si l'option est activée dans l'onglet « Graphiques » de la boîte de dialogue.

Bilan de l'optimisation : dans ces tableaux sont affichés les détails de l'optimisation pour la construction du plan d'expérience. Ce bilan présente le nombre d'itérations nécessaires à la réalisation du plan ainsi que la D-efficacité ou la diagonalité. Dans le cas de plans simples, la D-éfficacité est affichée, c'est un indicateur relatif d'efficacité, permettant la comparaison avec d'autres plans de même taille, le but est de maximiser cet indicateur qui est entre 0 et 1. La diagonalité est affichée lorsque l'on souhaite construire des plans multiples, c'est une mesure qui permet de mesurer que la confusion entre facteurs et interactions est minimale. Cette mesure est comprise entre 0 et 1, plus elle est proche de 1, plus la confusion est faible, ce qui est recherché. Le plan sélectionné au final est mis en gras dans le tableau.

Feuilles _Res : ces feuilles sont les feuilles qui s'affichent lorsque l'option « afficher les feuilles individuelles » est activée. Chaque feuille correspond à un questionnaire avec le nom de

l'analyse, le numéro du répondant ainsi qu'un tableau pour chaque choix. La dernière ligne de chaque tableau correspond au code à inscrire dans la cellule en dessous du tableau. Cette cellule doit être remplie par le répondant avec le code correspondant à son choix.

Exemple

Un exemple d'analyse conjointe basée sur le choix est disponible sur le Centre d'aide XLSTAT :

http://www.xlstat.com/demo-cbcf.htm

Bibliographie

Green, P.E. and Srinivasan, V. (1990). Conjoint analysis in Marketing: New Developments with implication for research and practice, Journal of Marketing, **54** (4), 3-19.

Gustafson, A., Herrmann, A. and Huber F. (eds.) (2001). Conjoint Measurement. Method and Applications, Springer.

Analyse conjointe

Utilisez cet outil pour effectuer une analyse conjointe basée sur les profils complets. Cet outil est inclus dans le module XLSTAT-Conjoint et doit être utilisé sur des plans d'expériences générés à l'aide de l'outil permettant de générer des plans pour l'analyse conjointe de XLSTAT-Conjoint.

Dans cette section :

Description

L'analyse conjointe est un processus complet d'analyse servant à l'analyse de nouveaux produits dans un contexte concurrentiel.

Cet outil permet de mener à bien l'étape consistant en l'analyse des résultats obtenus après le recueil des réponses auprès d'un échantillon d'individus.

Il s'agit dans le cas de profils complets de noter ou de classer un ensemble de produits générés à l'aide de plans d'expériences représentant des produits existants ou virtuels.

L'analyse se fait en utilisant deux types de méthodes statistiques :

L'analyse de la variance basée sur les moindres carrés (OLS).

L'analyse de la variance monotone (Kruskal, 1964) qui utilise des transformations monotones des scores ou des classements afin de mieux ajuster l'analyse de la variance (MONANOVA).

Ces deux approches sont décrites en détail dans les chapitres « Analyse de la variance » et « La régression monotone (MONANOVA) » de l'aide d'XLSTAT.

L'analyse conjointe permet donc d'obtenir pour chaque individu des utilités partielles associées à chaque modalité de chaque variable. Ces utilités brutes donnent une idée de l'impact de chaque modalité sur le processus de choix d'un produit.

En plus des utilités, l'analyse conjointe permet d'obtenir des importances associées à chaque variable et ce qui permettra de visualiser l'importance de chacune des variables dans le processus de choix associé à chaque individu.

L'analyse conjointe basée sur les profils complets détaille les résultats pour chaque individu séparément ce qui permet de conserver l'hétérogénéité des résultats. XLSTAT-Conjoint propose

aussi d'effectuer des classifications sur les individus. Ainsi, en utilisant les utilités obtenues, XLSTAT-Conjoint va obtenir des classes d'individus qui pourront être analysées et permettre de plus amples recherches. Les méthodes de classification utilisées dans XLSTAT- Conjoint sont la classification ascendante hiérarchique (voir le chapitre sur ce sujet dans l'aide de XLSTAT) et la méthode des k-means (voir le chapitre sur ce sujet dans l'aide de XLSTAT).

Type de données

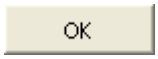
XLSTAT-Conjoint propose deux types de données en entrée de l'analyse conjointe : les classements et les notations. Il faut bien indiquer le type de données que vous utilisez car le traitement utilisé diffère légèrement.

En effet, avec un classement, le meilleur profil aura la valeur la plus basse alors qu'avec une notation, il aura la note la plus élevée.

Si l'option classement est sélectionnée, XLSTAT-Conjoint transforme les réponses de manière à inverser le classement et ceci permet d'interpréter de manière simple les utilités obtenues.

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.

Onglet **Général** :

Chargement automatique des données  : L'analyse conjointe nécessite le chargement de deux tableaux de données : un tableau concernant les réponses des individus et un tableau contenant les différents profils (produits évalués par les individus). Si les réponses des individus sont rassemblées sur une feuille contenant un plan d'expérience pour l'analyse conjointe généré par XLSTAT, vous pouvez charger les deux tableaux de données automatiquement.

Pour cela, il vous suffit de cliquer sur le bouton « baguette magique » puis de sélectionner n'importe quelle cellule de la feuille de résultat contenant le plan pour l'analyse conjointe généré par XLSTAT. Afin que les données soit chargées correctement il est important que vous n'ayez pas modifié manuellement la feuille de résultats contenant le plan généré par XLSTAT (pas d'ajouts de lignes ou de colonnes,…). Vous pouvez également charger les données en sélectionnant les différents tableaux séparément.

Réponses : sélectionnez les réponses des individus sous forme de notes ou de classements sur une échelle commune pour tous les individus. Si des en- têtes ont été sélectionnés, veuillez vérifier que l'option « Libellés des variables » est activée. Il s'agit de la partie de droite du tableau « plan d'analyse conjointe ».

Type de réponse : sélectionnez le type de réponses données par les individus (classement ou notation). Si les profils ont été classés alors l'ordre de classement est automatiquement inversé par XLSTAT-Conjoint afin de considérer que l'impact de la 1ère position est le plus fort et que celui de la dernière le moins fort.

Profils : sélectionnez les profils générés par l'outil de génération de plans d'expériences pour l'analyse conjointe de XLSTAT-Conjoint. Si des en- têtes ont été sélectionnés, veuillez vérifier que l'option « Libellés des variables » est activée. Il s'agit de la partie de gauche du tableau « plan d'analyse conjointe ». La 1ère colonne comprenant les numéros des profils ne doit pas être sélectionnée.

Libellés des variables : activez cette option si la première ligne des sélections (données, autre groupe) contient un libellé.

Poids des profils : activez cette option si vous voulez associer des poids aux profils. Sélectionnez les poids dans un vecteur vertical dont la taille est égal au nombre de profils. Si des en-têtes ont été sélectionnés, veuillez vérifier que l'option « Libellés des variables » est activée.

Cas d'exclusion : entrez la colonne des cas d'exclusion. Les cas d'exclusion sont des cas évalués par les répondants, mais qui ne sont pas incluses dans l'analyse conjointe ensuite. Ces cas sont intéressant à ajouter à votre modèle car ils vont permettre de vérifier sa validité. Dans le cas de l'analyse conjointe, une mesure du Tau de Kendall par répondants sera calculée pour ces cas spécifiquement.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Onglet **Options**:

Méthode : sélectionner la méthode que vous désirez utiliser.

Interactions / Niveau : activez cette option pour inclure des interactions dans le modèle puis entrez le niveau maximum d'interaction (valeur comprise entre 1 et 4).

Tolérance : activez cette option pour permettre à l'algorithme de calcul de ne pas prendre en compte les variables qui seraient soit constantes soit trop corrélées avec d'autres variables déjà utilisées dans le modèle (0.0001 par défaut).

Intervalle de confiance (%) : entrez l'étendue en pourcentage de l'intervalle de confiance à utiliser pour les différents tests, et pour le calcul des intervalles de confiance autour des paramètres et des prédictions. Valeur par défaut : 95%.

Contraintes : différentes contraintes sont disponibles pour les facteurs du modèle.

a1 = 0 : choisissez cette option pour que le paramètre de la première modalité de chaque facteur soit fixé à 0.

an = 0 : choisissez cette option pour que le paramètre de la dernière modalité de chaque facteur soit fixé à 0.

Somme (ai) = 0 : pour chaque facteur la somme des paramètres associés aux différentes modalités vaut 0. C'est cette option qui est activée par défaut en analyse conjointe.

Segmentation : activez cette option si vous désirez appliquer une méthode de classification sur les utilités partielles obtenues. Deux méthodes sont alors disponibles : la classification ascendante hiérarchique et la méthode des k-means.

Nombre de classes : dans le cas de la méthode des k-means, entrez le nombre de classes qui doivent être créées par l'algorithme.

Troncature : dans le cas de la classification ascendante hiérarchique, activez cette option si vous voulez que XLSTAT définisse **automatiquement** une troncature, et donc le nombre de classes à retenir, ou si vous voulez définir vous-même le **nombre de classes** à créer, ou le **niveau** auquel le dendrogramme doit être tronqué.

Conditions d'arrêt : le nombre d'itérations et l'indice de convergence pour l'algorithme MONANOVA peuvent être réglés.

Onglet **Données manquantes** :

Supprimer les réponses du répondant : activez cette option pour supprimer l'ensemble des réponses de chaque répondant comportant une ou plusieurs données manquantes.

Estimer les données manquantes : activez cette option pour estimer les données manquantes avant le début des calculs.

- **Moyenne ou mode** : activez cette option pour estimer les données manquantes en utilisant la moyenne des notations des individus pour le profil comportant des données manquantes.
- **Plus proche voisin** : activez cette option pour estimer les données manquantes d'un répondant en utilisant les notations du répondant le plus proche pour le profil correspondant.

Onglet **Sorties** :

Détails pour chaque individu : activez cette option pour afficher en plus des utilités partielles et des importances, l'ensemble des sorties qui suivent pour chaque individu.

Coefficients d'ajustement : activez cette option pour afficher le tableau avec les coefficients d'ajustement calculés pour chaque individu.

Analyse de la variance : activez cette option pour afficher le tableau d'analyse de la variance.

Type III SS : activez cette option pour afficher le tableau de l'analyse de la variance de Type III (*Type III Sum of Squares*).

Coefficients normalisés : activez cette option pour afficher les paramètres normalisés du modèle (coefficients bêta).

Prédictions et résidus : activez cette option pour afficher les prédictions et les résidus pour l'ensemble des observations.

Résultats par classe : si une méthode de classification a été sélectionnée, activez cette option pour afficher les résultats par classe.

Résultats par observations : si une méthode de classification a été sélectionnée, activez cette option pour afficher les résultats par individu.

Onglet Graphiques :

Graphiques des utilités : activez cette option pour afficher sur un graphique les utilités moyennes des différentes modalités.

Graphiques des importances : activez cette option pour afficher sur un graphique les importances moyennes des différentes variables.

Graphiques de régression : activez cette option pour afficher les graphiques de régression :

- **Coefficients normalisés** : activez cette option pour afficher sur un graphique les paramètres normalisés du modèle avec leur intervalle de confiance.

Graphiques de transformation : dans le cas de la méthode MONANOVA, activez cette option pour afficher les graphiques de transformation des données.

Dendrogramme : dans le cas de la classification ascendante hiérarchique, activez cette option pour afficher le dendrogramme.

- **Horizontal** : choisissez cette option pour afficher un dendrogramme horizontal.
- **Vertical** : choisissez cette option pour afficher un dendrogramme vertical.
- **Complet** : activez cette option pour afficher le dendrogramme complet (tous les objets sont représentés).
- **Tronqué** : activez cette option pour afficher le dendrogramme tronqué (le dendrogramme commence au niveau de la troncature).
- **Etiquettes** : activez cette option pour afficher les libellés des objets (dendrogramme complet) ou des classes (dendrogramme tronqué) sur le dendrogramme.
- **Couleurs** : activez cette option pour utiliser des couleurs pour représenter les différents groupes sur le dendrogramme complet.

Résultats

Informations sur les variables : dans ce tableau sont récapitulées toutes les informations sur les facteurs sélectionnés.

Utilités (Données individuelles) : dans ce tableau sont affichées les utilités associées à chaque modalité de chaque variable pour chacun des individus interrogés.

Tableau des écarts-types : Ce tableau rassemble les écarts-types associés aux utilités de chaque individu ainsi que l'erreur du modèle. Il est utile lors de l'application de la méthode de Simulation RFC-Bolse (cf. chapitre sur la simulation pour l'analyse conjointe).

Utilités (statistiques descriptives) : dans ce tableau sont affichées les statistiques descriptives associées aux utilités (pour tous les individus) avec l'utilité minimale, maximale, moyenne ainsi que l'écart-type.

Importances (Données individuelles) : dans ce tableau sont affichées les importances associées à chaque variable pour chacun des individus interrogés.

Importances (statistiques descriptives) : dans ce tableau sont affichées les statistiques descriptives associées aux importances (pour tous les individus) avec l'importance minimale, maximale, moyenne ainsi que l'écart- type.

Coefficients d'ajustement : dans ce tableau sont affichées les statistiques relatives à l'ajustement du modèle de régression :

- **Observations** : le nombre d'observations prises en compte dans les calculs. Dans les formules présentées ci-dessous n désigne le nombre d'observations.
- **Somme des poids** : la somme des poids des observations prises en compte dans les calculs. Dans les formules présentées ci-dessous W désigne la somme des poids.

- **DDL** : le nombre de degrés de liberté pour le modèle retenu (correspondant à la partie erreurs).
- R^2 : le coefficient de détermination du modèle. Ce coefficient, dont la valeur est comprise entre 0 et 1, n'est affiché que si la constante du modèle n'est pas fixée par l'utilisateur. Sa valeur est définie par

$$R^2 = 1 - \frac{\sum_{i=1}^{n} w_i (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} w_i (y_i - \bar{y}_i)^2}, \text{ avec } \bar{y} = \frac{1}{n} \sum_{i=1}^{n} w_i y_i$$

- Le R^2 s'interprète comme la proportion de la variabilité de la variable dépendante expliquée par le modèle. Plus le R^2 est proche de 1, meilleur est le modèle. L'inconvénient du R^2 est qu'il ne prend pas en compte le nombre de variables utilisées pour ajuster le modèle.
- R^2 **ajusté** : le coefficient de détermination ajusté du modèle. Le R^2 ajusté peut être négatif si le R^2 est voisin de zéro. Ce coefficient n'est affiché que si la constante du modèle n'est pas fixée par l'utilisateur. Sa valeur est définie par

$$\hat{R}^2 = 1 - (1 - R^2) \frac{W - 1}{W - p - 1}$$

Le R^2 ajusté est une correction du R^2 qui permet de prendre en compte le nombre de variables utilisées dans le modèle.

- **MCE** : la moyenne des carrés des erreurs (MCE) est définie par :

$$MCE = \frac{1}{W - p^*} \sum_{i=1}^{n} w_i (y_i - \hat{y}_i)^2$$

- **RMCE** : la racine de la moyenne des carrés des erreurs (RMCE) est la racine carrée de la MCE.
- **MAPE** : la *Mean Absolute Percentage Error* est calculée comme suit :

$$MAPE = \frac{100}{W} \sum_{i=1}^{n} w_i \left| \frac{y_i - \hat{y}_i}{y_i} \right|$$

- **DW** : le coefficient de Durbin-Watson est défini par

$$DW = \frac{\sum_{i=2}^{n} [(y_i - \hat{y}_i) - (y_{i-1} - \hat{y}_{i-1})]^2}{\sum_{i=1}^{n} w_i (y_i - \hat{y}_i)^2}$$

Ce coefficient correspond au coefficient d'autocorrélation d'ordre 1 et permet de vérifier que les résidus du modèle ne sont pas autocorrélés, sachant que l'indépendance des résidus est l'une des hypothèses de base de la régression linéaire. L'utilisateur pourra se référer à une table des coefficients de Durbin-Watson pour vérifier si l'hypothèse d'indépendance des résidus est acceptable.

- Cp : le coefficient Cp de Mallows est défini par

$$Cp = \frac{SCE}{\hat{\sigma}} + 2p^* - W$$

où SCE est la somme du carré des erreurs pour le modèle avec p variables explicatives, et où $\hat{\sigma}$ correspond à l'estimateur de la variance des résidus pour le modèle comprenant toutes les variables explicatives. Plus le coefficient Cp est proche de $p*$ moins le modèle est biaisé.

- **AIC** : le critère d'information d'Akaike (Akaike's Information Criterion) est défini par

$$AIC = W \ln\left(\frac{SCE}{W}\right) + 2p^*$$

Ce critère proposé par Akaike (1973) dérive de la théorie de l'information, et s'appuie sur la mesure de Kullback et Leibler (1951). C'est un critère de sélection de modèles qui pénalise les modèles pour lesquels l'ajout de nouvelles variables explicatives n'apporte pas suffisamment d'information au modèle, l'information étant mesurée au travers de la SCE. On cherche à minimiser le critère AIC.

- **SBC** : le critère bayésien de Schwarz (Schwarz's Bayesian Criterion) est défini par

$$SBC = W \ln\left(\frac{SCE}{W}\right) + \ln(W)p^*$$

Ce critère proposé par Schwarz (1978) est proche du critère AIC, et comme ce dernier on cherche à le minimiser.

- **PC** : le critère de prédiction d'Amemiya (Amemiya's Prediction Criterion) est défini par

$$PC = \frac{(1-R^2)(W+p^*)}{W-p^*}$$

Ce critère proposé par Amemiya (1980) permet comme le R^2 ajusté de tenir compte de la parcimonie du modèle.

- **Press** : la statistique du Press (predicted residual error sum of squares) n'est affichée que si l'option correspondante a été activée dans la boîte de dialogue. Elle est définie par

$$Press = \sum_{i=1}^{n} w_i (y_i - \hat{y}_{i(-i)})^2$$

où $\hat{y}_{i(-i)}$ est la prédiction pour l'observation i lorsque cette dernière n'est pas utilisée pour l'estimation des paramètres. On obtient alors

$$\text{Press RMCE } = \sqrt{\frac{\text{Press}}{W-p^*}}$$

Le Press RMCE peut alors être comparé au RMCE. Une différence importante entre les deux indique que le modèle est sensible à la présence ou absence de certaines observations dans le

modèle.

Si l'option d'affichage pour chaque individu est activée, les tableaux suivants sont affichés pour chaque individu indépendamment.

Coefficients d'ajustement (MONANOVA) : dans ce tableau sont affichées les statistiques relatives à l'ajustement du modèle de régression spécifique au cas de l'ANOVA monotone. Ces indices sont le lambda de Wilks, la trace de Pillai, le trace de Hotelling-Lawlet et la plus grande racine de Roy. Pour plus de détails sur ces statistiques, on peut voir l'aide sur la régression monotone (MONANOVA).

Le **tableau d'analyse de la variance** permet d'évaluer le pouvoir explicatif des variables explicatives. Le pouvoir explicatif est évalué en comparant l'ajustement (au sens des moindres carrés) du modèle final avec l'ajustement du modèle rudimentaire composé d'une constante égale à la moyenne de la variable dépendante.

Le tableau des **Type III SS** permet de visualiser l'influence du retrait d'une variable explicative sur l'ajustement du modèle, toutes les autres variables étant conservées, au sens de la somme des carrés des erreurs (SCE), de la moyenne des carrés des erreurs (MCE), du F de Fisher, ou de la probabilité associée au F de Fisher. Plus la probabilité est faible, plus la contribution de la variable au modèle est importante, toutes les autres variables étant déjà dans le modèle. Remarque : contrairement au cas des Type I SS, l'ordre de sélection des variables dans le modèle n'influe pas sur les valeurs obtenues.

Le tableau des **coefficients normalisés** (aussi appelés coefficients bêta) permet de comparer le poids relatif des variables. Plus la valeur absolue d'un coefficient est élevée, plus le poids de la variable correspondante est important. Lorsque l'intervalle de confiance autour des coefficients normalisés comprend la valeur 0 (cela est facilement visible sur le graphique des coefficients normalisés), le poids d'une variable dans le modèle n'est pas significatif.

Dans le tableau des **prédictions et résidus** sont donnés pour chaque observation, son poids, la valeur observée de la variable dépendante, la valeur de la variable dépendante transformée, la prédiction du modèle, les résidus et les intervalles de confiance. Deux types d'intervalles de confiance sont affichés : un intervalle de confiance autour de la moyenne (correspondant au cas où l'on ferait la prédiction pour un nombre infini d'observations avec un ensemble de valeurs données des variables explicatives) et un intervalle autour de la prédiction ponctuelle (correspondant au cas d'une prédiction isolée pour des valeurs données des variables explicatives). Le second intervalle est toujours plus grand que le premier, les aléas étant plus importants.

Le **graphique** qui suit permet de visualiser les transformations monotones obtenues par l'algorithme MONANOVA.

Si une méthode de classification a été appliquée, les résultats suivant sont aussi affichés :

Barycentre des classes : dans ce tableau sont affichées les valeurs moyennes des utilités des différentes variables par classe. Cela permet de mieux identifier les profils d'individus des différentes classes.

Matrice de proximité (dans le cas d'une CAH) : cette matrice contient les proximités entres les individus calculées sur leurs utilités associées.

Dendrogrammes (dans le cas d'une CAH) : le dendrogramme complet permet de visualiser le regroupement progressif des objets. Si une troncature a été demandée, un trait en pointillé marque le niveau auquel est effectuée la troncature. Le dendrogramme tronqué permet de visualiser les classes après la troncature

Résultats par classe : les statistiques descriptives des classes (nombre d'objets, somme des poids, variance intra-classe, distance minimale au barycentre, distance maximale au barycentre, distance moyenne au barycentre) sont affichées dans la première partie du tableau. Les objets sont affichés dans la seconde partie.

Résultats par objet : dans ce tableau est indiquée pour chaque objet sa classe d'affectation dans l'ordre initial des objets.

Exemple

Un exemple d'analyse conjointe est disponible sur le Centre d'aide XLSTAT :

http://www.xlstat.com/demo-conjointf.htm

Bibliographie

Green, P.E. and Srinivasan, V. (1990). Conjoint analysis in Marketing: New Developments with implication for research and practice, Journal of Marketing, **54** (4), 3-19.

Gustafson, A., Herrmann, A. and Huber F. (eds.) (2001). Conjoint Measurement. Method and Applications, Springer.

Guyon, H. and Petiot J.-F. (2011) Market share predictions: a new model with rating-based conjoint analysis. *International Journal of Market Research,* **53(6)**, 831-857.

Analyse conjointe basée sur le choix

Utilisez cet outil pour effectuer une analyse conjointe basée sur le choix. Cet outil est inclus dans le module XLSTAT-Conjoint et doit être utilisé sur des plans d'expérience générés à l'aide de l'outil permettant de générer des plans pour l'analyse conjointe basée sur le choix de XLSTAT-Conjoint.

Dans cette section :

Description

L'analyse conjointe est un processus complet d'analyse servant à l'analyse de nouveaux produits dans un contexte concurrentiel.

Cet outil permet de mener à bien l'étape consistant en l'analyse des résultats obtenus après le recueil des réponses obtenues auprès d'un échantillon de personnes.

Il s'agit dans le cas de modèles de choix de choisir entre plusieurs profils proposés sous forme d'un choix à un ensemble d'individus. Ainsi, un certain nombre de choix sont donnés à tous les individus (on sélectionnera un produit parmi plusieurs produits générés). Une option de non-choix est aussi disponible.

L'analyse de ces choix peut se faire en utilisant :

- - un modèle logit multinomial spécifique basé sur le modèle logit conditionnel. Pour plus de détails, voir l'aide sur le modèle logit conditionnel. Dans ce cas, on obtient des utilités agrégées, c'est-à-dire une utilité pour chaque modalité de chaque variable associée à tous les individus. Il est alors impossible d'effectuer des classifications sur les individus.

- - un algorithme bayésien hiérarchique qui permet d'obtenir des résultats individu par individu. Les différents paramètres sont estimés au niveau individuel via une procédure itérative (échantillonnage de Gibbs) qui tient compte à la fois du choix de chacun des individus ainsi que de la distribution globale de ces choix. Les estimations au niveau individuel permettent d'améliorer la précision des simulations de marché ainsi qu'une meilleure compréhension de la structure du marché.

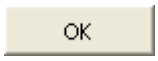
XLSTAT-Conjoint propose d'inclure une variable de segmentation dans le cas du modèle classique qui permettra de construire des modèles distincts suivant les modalités de cette

variable de segmentation et d'effectuer des classifications sur les individus comme dans le cas de l'analyse conjointe classique lorsqu'on utilise une estimation par l'algorithme bayésien hiérarchique.

En plus des utilités, l'analyse conjointe permet d'obtenir des importances associées à chaque variable et qui permettent de visualiser l'importance de chacune des variables dans le processus de choix.

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.

Onglet **Général** :

Chargement automatique des données  : L'analyse conjointe basée sur le choix nécessite le chargement de trois tableaux de données : un tableau concernant les réponses des individus, un tableau concernat les différents choix proposés et un tableau contenant les différents profils (produits à choisir par les individus). Si les réponses des individus sont rassemblées sur une feuille contenant un plan d'expérience pour l'analyse conjointe basée sur le choix généré par XLSTAT, vous pouvez charger les trois tableaux de données automatiquement. Pour cela, il vous suffit de cliquer sur le bouton « baguette magique » puis de sélectionner n'importe quelle cellule de la feuille de résultat contenant le plan pour l'analyse conjointe basée sur le choix généré par XLSTAT. Afin que les données soit chargées correctement il est important que vous n'ayez pas modifié manuellement la feuille de résultats contenant le plan généré par XLSTAT (pas d'ajouts de lignes ou de colonnes,…). Vous pouvez également charger les données en sélectionnant les différents tableaux séparément.

Réponses : sélectionnez les réponses des individus sous forme de nombres associés aux choix effectués. Ainsi s'il y a 3 possibilités, il faudra entrer soit 1, soit 2, soit 3. Si l'individu n'a rien choisi, un 0 doit être ajouté. Si des en-têtes ont été sélectionnés, veuillez vérifier que l'option « Libellés des variables » est activée. Cette sélection correspond aux choix effectués par les individus, elle correspond à la partie droite du tableau « Plan d'analyse conjointe ».

Choix : sélectionnez les choix générés par l'outil de génération d'analyses conjointes basées sur le choix d'XLSTAT-Conjoint. Si des en-têtes ont été sélectionnés, veuillez vérifier que l'option « Libellés des variables » est activée. La 1ère colonne comprenant les numéros des sélections ne doit pas être sélectionnée.

Profils : sélectionnez les profils générés par l'outil de génération d'analyses conjointes basées sur le choix d'XLSTAT-Conjoint. Si des en-têtes ont été sélectionnés, veuillez vérifier que l'option « Libellés des variables » est activée. La 1ère colonne comprenant les numéros des profils ne doit pas être sélectionnée. Cette sélection correspond au tableau des profils.

Poids des réponses : activez cette option si vous voulez associer des poids aux réponses des individus. Sélectionnez les poids dans un vecteur vertical dont la taille est égale au nombre d'individus. Si des en-têtes ont été sélectionnés, veuillez vérifier que l'option « Libellés des variables » est activée.

Variable de segmentation : activez cette option si vous voulez associer une variable de groupes aux individus (variable de segmentation qualitative). Sélectionnez les groupes dans un vecteur vertical dont la taille est égal au nombre d'individus. Si des en-têtes ont été sélectionnés, veuillez vérifier que l'option « Libellés des variables » est activée.

Cas d'exclusion : entrez la colonne des cas d'exclusion. Les cas d'exclusion sont des cas évalués par les répondants, mais qui ne sont pas incluses dans l'analyse conjointe ensuite. Ces cas sont intéressant à ajouter à votre modèle car ils vont permettre de vérifier sa validité. Dans le cas du modèle logit multinomial, une mesure du rlh sera calculée pour ces cas spécifiquement. Dans le cas du modèle bayésien hiérarchique, une mesure du rlh par répondents sera calculée pour ces cas.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Onglet **Options** :

Méthode : sélectionnez la méthode que vous désirez utiliser.

Interactions / Niveau : activez cette option pour inclure des interactions dans le modèle puis entrez le niveau maximum d'interaction (valeur comprise entre 1 et 4).

Tolérance : activez cette option pour permettre à l'algorithme de calcul de la régression OLS de ne pas prendre en compte les variables qui seraient soit constantes soit trop corrélées avec d'autres variables déjà utilisées dans le modèle (0.0001 par défaut).

Intervalle de confiance (%) : entrez l'étendue en pourcentage de l'intervalle de confiance à utiliser pour les différents tests, et pour le calcul des intervalles de confiance autour des paramètres et des prédictions. Valeur par défaut : 95%.

Contraintes : différentes contraintes sont disponibles pour les facteurs du modèle.

a1 = 0 : choisissez cette option pour que le paramètre de la première modalité de chaque facteur soit fixé à 0.

an = 0 : choisissez cette option pour que le paramètre de la dernière modalité de chaque facteur soit fixé à 0.

Somme (ai) = 0 : pour chaque facteur la somme des paramètres associés aux différentes modalités vaut 0.

Options du bayésien (uniquement en CBC/HB) : le nombre d'itérations pour la période de chauffe et le temps maximal de l'algorithme peuvent être réglés.

Segmentation (uniquement en CBC/HB) : activez cette option si vous désirez appliquer une méthode de classification sur les utilités partielles obtenues. Deux méthodes sont alors disponibles : la classification ascendante hiérarchique et la méthode des k-means.

- **Nombre de classes** : dans le cas de la méthode des k-means, entrez le nombre de classes qui doivent être créées par l'algorithme.
- **Troncature** : dans le cas de la classification ascendante hiérarchique, activez cette option si vous voulez que XLSTAT définisse automatiquement une troncature, et donc le nombre de classes à retenir, ou si vous voulez définir vous-même le nombre de classes à créer, ou le niveau auquel le dendrogramme doit être tronqué.

Conditions d'arrêt : le nombre d'itérations et l'indice de convergence pour l'algorithme de Newton-Raphson peuvent être réglés.

Onglet **Données manquantes** :

Supprimer les réponses du répondant : activez cette option pour supprimer l'ensemble des réponses des répondants possédant une ou plusieurs valeurs manquantes.

Onglet **Sorties** :

Coefficients d'ajustement : activez cette option pour afficher le tableau avec les coefficients d'ajustement calculés pour chaque individu.

Analyse de Type III : activez cette option pour afficher le tableau de l'analyse de la variance de Type III (*Type III Sum of Squares*).

Coefficients normalisés : activez cette option pour afficher les paramètres normalisés du modèle (coefficients bêta).

Détails pour les observations (uniquement en CBC/HB) : activez cette option pour afficher les caractéristiques de la distribution *a posteriori* pour chaque individu dans le cas bayésien hiérarchique.

Onglet Graphiques :

Graphiques des utilités : activez cette option pour afficher sur un graphique les utilités moyennes des différentes modalités.

Graphiques des importances : activez cette option pour afficher sur un graphique les importances moyennes des différentes variables.

Dendrogramme (uniquement en CBC/HB) : dans le cas de la classification ascendante hiérarchique, activez cette option pour afficher le dendrogramme.

- **Horizontal** : choisissez cette option pour afficher un dendrogramme horizontal.
- **Vertical** : choisissez cette option pour afficher un dendrogramme vertical.
- **Complet** : activez cette option pour afficher le dendrogramme complet (tous les objets sont représentés).
- **Tronqué** : activez cette option pour afficher le dendrogramme tronqué (le dendrogramme commence au niveau de la troncature).
- **Etiquettes** : activez cette option pour afficher les libellés des objets (dendrogramme complet) ou des classes (dendrogramme tronqué) sur le dendrogramme.
- **Couleurs** : activez cette option pour utiliser des couleurs pour représenter les différents groupes sur le dendrogramme complet.

Résultats

Informations sur les variables : dans ce tableau sont récapitulées toutes les informations sur les facteurs sélectionnés.

Utilités : dans ce tableau sont affichées les utilités associées à chaque modalité de chaque variable ainsi que l'écart-type associé.

Importances : dans ce tableau sont affichées les importances associées à chaque variable.

Coefficients d'ajustement : dans ce tableau est affichée une série de statistiques pour le modèle indépendant (correspondant au cas où la combinaison linéaire des variables explicatives se réduit à une constante) et pour le modèle ajusté.

- **Observations** : le nombre total d'observations prises en compte (somme des poids des observations) ;
- **Somme des poids** : le nombre total d'observations prises en compte (somme des poids des observations multipliés par les poids dans la régression) ;
- **DDL** : degrés de liberté ;
- **-2 Log(Vrais.)** : le logarithme de la fonction de vraisemblance associée au modèle;
- R^2**(McFadden) :** coefficient compris comme le R^2 entre 0 et 1 qui mesure le bon ajustement du modèle. Ce coefficient est égal à 1 moins le rapport de la vraisemblance du modèle ajusté sur la vraisemblance du modèle indépendant ;
- R^2**(Cox et Snell) :** coefficient compris comme le R^2 entre 0 et 1 qui mesure le bon ajustement du modèle. Ce coefficient est égal à 1 moins le rapport de la vraisemblance du modèle ajusté sur la vraisemblance du modèle indépendant, le rapport étant porté à l'exposant $\frac{2}{S_w}$, où S_w est la somme des poids ;
- R^2**(Nagelkerke) :** coefficient compris comme le R^2 entre 0 et 1 qui mesure le bon ajustement du modèle. Ce coefficient est égal au rapport du R^2 de Cox et Snell, divisé par 1 moins le la vraisemblance du modèle indépendant portée à l'exposant $\frac{2}{S_w}$;
- **AIC** : le critère d'information d'Akaike (Akaike's Information Criterion) ;
- **SBC** : le critère bayésien de Schwarz (Schwarz's Bayesian Criterion).
- **Itérations** : nombre d'itérations avant convergence.
- **rlh** : racine de la vraisemblance. Cette valeur varie entre 0 et 1, 1 correspond à un modèle qui ajuste parfaitement les données.
- **rlh par individu** : La valeur RLH (Root LikeliHood) est un indice de 0 à 1. Plus la valeur RLH d'un répondant est élevée, plus le répondant a répondu de manière cohérente aux questions de choix.

Coefficients d'ajustement (logit conditionnel) : dans ce tableau sont affichées les statistiques relatives à l'ajustement du modèle logit conditionnel (rapport de vraisemblance, borne supérieur du rapport de vraisemblance, indice d'Aldrich-Nelson, indice de Cragg-Uhler 1 et 2, indice d'Estrella, indice d'Estrella ajusté et indice de Veal-Zimmermann). Voir l'aide sur le modèle logit conditionnel pour plus de détails.

Test de l'hypothèse nulle H0 : **Y=p0** : l'hypothèse H_0 correspond au modèle indépendant qui donne la probabilité p_0 quelques soient les valeurs des variables explicatives ; on cherche à

vérifier si le modèle ajusté est significativement plus performant que ce modèle. Trois tests sont proposés : le test du rapport des vraisemblance (-2 Log(Vrais.)), le test du Score, et le test test de Wald. Les trois statistiques suivent une loi du χ^2 dont les degrés de liberté sont indiqués.

Le tableau des **coefficients normalisés** (aussi appelés coefficients bêta) permet de comparer le poids relatif des variables. Plus la valeur absolue d'un coefficient est élevée, plus le poids de la variable correspondante est important. Lorsque l'intervalle de confiance autour des coefficients normalisés comprend la valeur 0 (cela est facilement visible sur le graphique des coefficients normalisés), le poids d'une variable dans le modèle n'est pas significatif.

Exemple

Un exemple d'analyse conjointe basée sur le choix est disponible sur le Centre d'aide XLSTAT :

http://www.xlstat.com/demo-cbcf.htm

Bibliographie

Green, P.E. and Srinivasan, V. (1990). Conjoint analysis in Marketing: New Developments with implication for research and practice, Journal of Marketing, **54** (4), 3-19.

Gustafson, A., Herrmann, A. and Huber F. (eds.) (2001). Conjoint Measurement. Method and Applications, Springer.

Lenk P. J., DeSarbo W. S., Green P. E. and Young M. R. (1996). Hierarchical Bayes Conjoint Analysis: Recovery of Partworth Heterogeneity from Reduced Experimental Designs. *Marketing Science*, **15**, 173-191.

Générateur de marché

Utilisez cet outil pour générer un marché qui sera par la suite utilisé pour simuler les parts de marchés des différents produits à l'aide de l'outil de simulation.

Dans cette section :

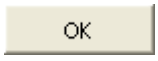
Description

Les résultats d'une analyse conjointe ou d'une analyse conjointe basée sur le choix peuvent être utilisés pour simuler les parts de marchés de différents produits à l'aide de l'outil de simulation. Pour cela il est nécessaire de générer un tableau décrivant les différents produits du marché à simuler. L'outil générateur de marché permet de créer ce tableau de marché avec les différents produits à partir des informations sur les variables obtenues dans les résultats de l'analyse conjointe ou de l'analyse conjointe basée sur le choix.

Une fois ces informations entrées dans la boîte de dialogue, il suffit de cliquer sur OK, et pour chaque attribut de chaque produit, il vous sera demandé de choisir la modalité à ajouter. Après que chaque produit a été construit, il vous est possible de sortir de l'outil afin de générer le tableau ou alors de continuer jusqu'au dernier produit.

Boîte de dialogue

La boîte de dialogue est composée d'un seul onglet correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.

Information sur les variables : sélectionnez le tableau (y compris les libellés des colonnes) intitulé « Information sur les variables » obtenus dans la feuille de résultat d'une analyse

conjointe ou d'une analyse conjointe basée sur le choix. Ce tableau comporte une colonne décrivant le nom des différentes variables (attributs), une colonne décrivant le nombre de modalités des différentes variables puis autant de colonnes que le nombre maximal de modalités.

Nombre de produits : entrez le nombre de produits que vous souhaitez générer.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Simulation pour l'analyse conjointe

Utilisez cet outil pour simuler des marchés en utilisant les utilités obtenues avec les outils d'analyse de XLSTAT-Conjoint.

Dans cette section :

Description

L'analyse conjointe est un processus complet d'analyse servant à l'analyse de nouveaux produits dans un contexte concurrentiel.

Une fois votre analyse conjointe menée à bien, l'intérêt majeur de cette approche est de pouvoir simuler des marchés réels ou virtuels en utilisant uniquement les informations obtenues par l'analyse. Ainsi, on pourra simuler un marché comprenant des produits qui n'ont pas été testés par les individus interrogés lors de l'analyse conjointe.

L'analyse conjointe (qu'elle soit basée sur les profils complets ou sur le choix) permet d'obtenir des utilités associées à chacune des modalités des facteurs d'intérêt. Ces utilités partielles vont permettre de calculer une utilité globale pour n'importe quel nouveau produit. Ces utilités permettent alors de calculer des parts de marché associées à chaque produit dans un marché « idéal ».

XLSTAT-Conjoint permet d'obtenir ces parts de marchés en utilisant de nombreuses méthodes (first choice, logit, Bradley-Terry-Luce, randomized first choice). Ces méthodes seront détaillées par la suite.

Les parts de marché obtenues peuvent être ensuite analysées afin d'évaluer la possible introduction d'un nouveau produit sur le marché. Les résultats de ces simulations restent néanmoins dépendants de la connaissance du marché réel et de la prise en compte de tous les facteurs important associé à chaque produit dans l'analyse conjointe.

XLSTAT-Conjoint permet aussi d'ajouter des poids aux modalités des variables ou aux individus. XLSTAT-Conjoint permet aussi de prendre en compte des groupes d'individus lorsqu'une variable de groupe (segmentation) est disponible. Elle peut être obtenue, par exemple, lors de la classification associée à l'analyse conjointe.

Type de données

XLSTAT-Conjoint propose deux modèles d'analyse conjointe. L'analyse en profils complets est basée sur des utilités avec une constante. On sélectionne donc les utilités et leur constante (sans la colonne avec les noms des modalités). Dans le cas de l'analyse conjointe basée sur le choix (CBC), il n'y a pas de constante et on sélectionne alors la colonne des utilités sans les libellés associés au nom des modalités.

Dans XLSTAT-Conjoint, il faut toujours sélectionner le tableau d'information sur les variables en entier. D'autre part, les produits à simuler peuvent être générés grâce à la fonction « Générateur de marché » en utilisant le tableau d'information sur les variables généré lors d'une analyse conjointe.

Méthode de simulation

XLSTAT-Conjoint propose 4 méthodes de simulation des parts de marché.

La première étape est constituée par le calcul des utilités associées à chaque nouveau produit. Ainsi pour une analyse conjointe basée sur le choix analysant des chaussures d'homme avec 3 variables : leur prix (50 euros, 100 euros, 150 euros), leur finition (tissu, cuir, daim) et leur couleur (marron, noir). On aura donc un tableau d'utilités partielles avec 8 lignes et une colonne.

On veut simuler un marché avec entre autres une chaussure à 100 euros, en cuir noir. L'utilité de ce produit est donc : $U_{P1} = U_{Prix-100} + U_{F-Cuir} + U_{C-Noir}$

On calcule cette utilité pour chaque produit du marché et on cherche la probabilité de choix de ce produit en utilisant l'une des différentes méthodes d'estimation :

- First choice : c'est la méthode la plus basique, on sélectionne le produit ayant l'utilité maximale avec une probabilité de 1.
- Logit : c'est une méthode basée sur la fonction exponentielle pour trouver la probabilité, elle est plus précise que la méthode first choice et lui est généralement préférée. Elle a l'inconvénient de supposer l'hypothèse IIA (hypothèse d'indépendance des alternatives non pertinentes). Elle se calcule pour le produit $P1$: $P_{P1} = \frac{\exp(U_{P1}\beta)}{sum_i \exp(U_{Pi}\beta)}$ avec $\beta = 1$ ou 2.
- Bradley-Terry-Luce : c'est une méthode proche de la méthode logit mais sans utilisation de la fonction exponentielle. Elle suppose toujours l'hypothèse IIA et demande des utilités positives (dans le cas $\beta = 1$). Elle se calcule pour le produit $P1$: $P_{P1} = \frac{U_{P1}^{\beta}}{\sum_i U_{Pi}^{\beta}}$ avec $\beta = 1$ ou 2.
- Randomized first choice : c'est une méthode à mi-chemin entre logit et first choice. Elle a l'avantage de ne pas supposer l'hypothèse IIA et se base sur un principe simple : on génère un très grand nombre de nombres issus d'une distribution de Gumbel et on crée une nouvelle série d'utilités en utilisant les utilités initiales auxquelles on ajoute les nombres générés. Pour chaque série d'utilités créées, on utilise la méthode first choice pour sélectionner l'un des produits. On va donc accepter de légères variations autour des

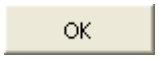
valeurs calculées des utilités. Cette méthode est la plus avancée mais aussi la plus adaptée au cas de l'analyse conjointe.

- RFC-Bolse : Dans le cas d'une analyse conjointe basée sur les profils complets, la méthode Randomized First Choice BOLSE (RFC-Bolse) a été présentée pour pallier aux problèmes de la méthode Randomized First Choice (RFC). En effet, la distribution associée à cette méthode (RFC) n'est pas adaptée au cas des profils complets avec des utilités individuelles. La distribution normale centrée est utilisée avec les écarts-types associés aux utilités dans le modèle de régression estimé. Comme la méthode RFC, la méthode RFC-Bolse ajoute une erreur aléatoire unique aux utilités et calcule ensuite les parts de marché. Pour chaque série d'utilités créées, on utilise la méthode first choice pour sélectionner l'un des produits. On va donc accepter de légères variations autour des valeurs calculées des utilités.

Lorsque l'on a plus d'une colonne d'utilités (cas d'une analyse conjointe en profils complets), on fera la moyenne des probabilités obtenues pour chaque individu.

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.

Onglet **Général** :

Chargement automatique des données  : La simulation pour l'analyse conjointe nécessite le chargement de trois tableaux de données : le tableau des utilités, le tableau d'information sur les variables généré lors d'une analyse conjointe avec XLSTAT et le tableau contenant le marché à simuler. Si vous utilisez les résultats d'une analyse conjointe réalisée avec XLSTAT et

que vous avez généré le marché avec la fonction XLSTAT « Générateur de marché », vous pouvez charger les différents tableaux de données automatiquement. Pour cela, il vous suffit de cliquer sur le bouton « baguette magique » puis de sélectionner n'importe quelle cellule de la feuille contenant les résultats de l'analyse conjointe générée par XLSTAT ainsi que n'importe quelle cellule de la feuille de résultat contenant le marché généré par XLSTAT. Afin que les données soit chargées correctement il est important que vous n'ayez pas modifié manuellement les feuilles de résultats générées par XLSTAT (pas d'ajouts de lignes ou de colonnes…). Vous pouvez également charger les données en sélectionnant les différents tableaux séparément.

Tableau des utilités : sélectionnez les utilités obtenues avec l'outil d'analyse conjointe (classique ou CBC). Si des en-têtes ont été sélectionnés, veuillez vérifier que l'option « Libellés des variables » est activée. Veillez à ne pas sélectionner les noms des modalités des variables.

Information sur les variables : sélectionnez le tableau « information sur les variables généré à l'aide de l'outil d'analyse conjointe de XLSTAT- Conjoint. Si des en-têtes ont été sélectionnés, veuillez vérifier que l'option « Libellés des variables » est activée.

Modèle d'analyse : sélectionnez le type de modèle d'analyse conjointe : les profils complets (analyse classique, dans ce cas les utilités ont aussi une constante) ou l'analyse basée sur le choix CBC (dans ce cas il n'y a pas de constante associée aux utilités).

Marché à simuler : sélectionnez les produits à simuler. Les produits seront répartis dans un tableau avec un produit par ligne et une variable par colonne. Si des en-têtes ont été sélectionnés, veuillez vérifier que l'option « Libellés des variables » est activée.

Identifiant produit : activez cette option si vous voulez associer des noms aux produits du marché à simuler. Sélectionnez les noms dans un vecteur vertical dont la taille est égal au nombre de produits. Si des en-têtes ont été sélectionnés, veuillez vérifier que l'option « Libellés des variables » est activée.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des variables : activez cette option si la première ligne des sélections (données, autre groupe) contient un libellé.

Poids des modalités : activez cette option si vous voulez associer des poids aux modalités des variables. Sélectionnez les poids dans un vecteur vertical dont la taille est égale au nombre total de modalités. Si des en- têtes ont été sélectionnés, veuillez vérifier que l'option « Libellés des variables » est activée.

Variable de groupe : activez cette option si vous voulez associer une variable de groupes aux individus. Cette variable pourra être vue comme une variable de segmentation. Sélectionnez les

groupes dans un vecteur vertical dont la taille est égal au nombre d'individus. Si des en-têtes ont été sélectionnés, veuillez vérifier que l'option « Libellés des variables » est activée.

Poids des réponses : activez cette option si vous voulez associer des poids aux réponses des individus. Sélectionnez les poids dans un vecteur vertical dont la taille est égal au nombre d'individus. Si des en-têtes ont été sélectionnés, veuillez vérifier que l'option « Libellés des variables » est activée.

Onglet **Options** :

Interactions / Niveau : activez cette option si des interactions ont été sélectionnées lors de l'analyse conjointe. Puis entrez le niveau maximum d'interaction (valeur comprise entre 1 et 3).

Nombre de simulations : entrez le nombre de simulations à générer pour la méthode « randomized first choice ».

Tableau des écarts-types : sélectionnez le tableau des écarts-types obtenus en sortie de l'analyse conjointe basée sur les profils complets. Si des en-têtes ont été sélectionnés, veuillez vérifier que l'option « Libellés des variables » est activée.

Onglet Graphiques :

Graphiques des parts de marché : activez cette option pour afficher les graphiques des parts de marché :

- **Diagrammes en secteur** : activez cette option pour afficher les diagrammes en secteur (camembert).
- **Comparer à l'échantillon total** : si des groupes d'individus ont été sélectionnés, activez cette option pour afficher des diagrammes superposés.

Résultats

Informations sur les variables : dans ce tableau sont récapitulées toutes les informations sur les facteurs sélectionnés. Si des interactions ont été sélectionnées, la liste des ces interactions est inscrite en dessous du tableau.

Marché à simuler : dans ce tableau sont affichées les produits utilisés pour effectuer la simulation. Habituellement, les premiers produits correspondent aux produits actuellement sur le marché, alors que le dernier produit correspond à un nouveau produit que l'on veut tester sur un marché existant.

Relancer l'analyse : afin de voir l'influence des caractéristiques d'un produit sur les différentes parts de marchés, vous pouvez modifier les modalités du dernier produit. Une fois ces modalités modifiées, en cliquant sur le bouton « Relancer l'analyse », les parts de marchés et les graphiques associés seront automatiquement mis à jour en fonction des nouvelles modalités choisies.

Parts de marché : dans ce tableau sont affichées les parts de marché associées à chaque produit. Si des groupes ont été sélectionnés, la première colonne représente le marché global

et les colonnes suivantes sont associées à chaque groupe.

Graphiques des parts de marché : le premier diagramme en secteur permet d'analyser la marché global. Si des groupes ont été sélectionnés, les diagrammes suivants sont associés aux différents groupes. Si l'option comparer à l'échantillon total est activée, les graphiques sont superposés avec en second plan les parts de marché sur le marché complet et au premier plan les parts de marché sur le groupe d'individus étudié.

Utilités / Parts de marché : dans ce tableau, qui n'apparaît que si aucun groupe n'est sélectionné, sont affichées les utilités calculées, les parts de marché ainsi que les écarts-type (quand cela est possible) associés à chaque produit du marché.

Parts de marché (individuelles) : dans ce tableau, qui n'apparaît que si aucun groupe n'est sélectionné et dans le cadre d'une analyse conjointe basée sur les profils complets, sont affichées les parts de marché obtenues à partir des utilités individuelles.

Exemple

Un exemple d'analyse conjointe basée sur les profils complets est disponible sur le Centre d'aide XLSTAT :

http://www.xlstat.com/demo-conjointf.htm

Un exemple d'analyse conjointe basée sur le choix est disponible sur leCentre d'aide XLSTAT :

http://www.xlstat.com/demo-cbcf.htm

Bibliographie

Green, P.E. and Srinivasan, V. (1990). Conjoint analysis in Marketing: New Developments with implication for research and practice. *Journal of Marketing*, **54(4)**, 3-19.

Gustafson, A., Herrmann, A. and Huber F. (eds.) (2001). Conjoint measurement. Method and applications, Springer.

Guyon, H. and Petiot J.-F. (2011) Market share predictions: a new model with rating-based conjoint analysis. *International Journal of Market Research,* **53(6)**, 831-857.

Plans d'expériences pour la méthode MaxDiff

Utilisez cet outil pour générer des plans d'expériences associés à la méthode MaxDiff afin de faire ressortir l'importance de certains attributs.

Dans cette section :

Description

La méthode MaxDiff est une méthode introduite par Jordan Louvière qui permet de connaître l'importance d'attributs en utilisant une présentation originale. On présente au répondant une série d'attributs parmi lesquels celui-ci doit choisir le plus important et le moins important.

Cette méthode nécessite deux étapes :

- Une étape de génération de plan d'expérience durant laquelle des combinaisons d'attributs sont construites afin d'être présentées aux répondants.
- Une fois les résultats précédents obtenus, une méthode de régression Bayésienne Hiérarchique est utilisée afin de faire ressortir les attributs importants de chaque individu interrogé. Un modèle logit conditionnel, peut également être utilisé, mais les résultats fournis ne concernent alors que les utilités des attributs.

Les groupes d'attributs à présenter sont générés en utilisant des plans d'expériences. Ce processus de création utilise une méthode de plans en blocs incomplets équilibrés (voir l'aide sur les plans d'expériences pour l'analyse sensorielle).

XLSTAT permet de générer plusieurs plans uniques, ce qui est un avantage notamment lorsque l'on souhaite interroger un grand nombre de personnes. Le nombre de combinaisons différentes étant plus important, les plans de l'analyse seront plus robustes dans l'analyse des effets. De plus, le fait d'inclure des plans différents réduit l'impact du contexte psychologique et des effets d'ordre.

Les attributs sont répartis dans des groupes de choix parmi lesquels le répondant va devoir choisir le plus important et le moins important (ou le meilleur et le pire).

La méthode MaxDiff (Maximum Difference Scaling)

La première étape de la méthode MaxDiff nécessite de choisir un certain nombre d'attributs décrivant un produit. Ainsi, par exemple, si on cherche à introduire un nouveau produit sur un marché, on pourra choisir comme attributs pouvant être importants : son prix, sa qualité, sa longévité… Cette première étape est primordiale et se fera à l'aide des experts du marché étudié.

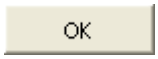
Une fois les attributs sélectionnés, il faudra générer le plan d'expérience et demander à chaque répondant de sélectionner l'attribut le plus important et l'attribut le moins important pour chaque choix.

Le nombre de choix et d'attributs par choix doit être choisi en fonction du nombre d'attributs. Tout en sachant que trop de choix peut être préjudiciable pour la qualité de réponse des individus interrogés.

XLSTAT-Conjoint permet d'obtenir un tableau global mais aussi des tableaux pour chaque individu séparément. Il suffira à celui-ci de remplir une case dans un tableau Excel afin que sa réponse soit comptabilisée dans l'analyse.

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.

Onglet **Général** :

Nom de l'analyse : entrez le nom de l'analyse que vous allez réaliser.

Attributs : sélectionnez le nombre d'attributs qui vont être testés lors de cette analyse MaxDiff.

Nombre de répondants : entrez le nombre de répondants dans votre analyse.

Nombre de comparaisons (combinaisons) : entrez le choix à effectuer pour chaque répondant.

Attributs (Choix) par comparaison (combinaison) : entrez le nombre d'attributs présenté lors de chaque choix.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Terminologie : choisissez parmi les possibilités offertes celle qui correspond le mieux au cadre de votre analyse.

Onglet **Sorties** :

Format des données : choisissez parmi l'un des deux formats dans lequel vous voulez que votre plan d'expérience soit affiché : Combinaison/Répondant ou Répondant/Combinaison.

Nombre de plans : activez cette option si vous souhaitez générer plusieurs plans.

- **Répondants par plan** : entrez un vecteur de la taille du nombre de plans, contenant le nombre de répondants par plan souhaité.
- **Comparaisons par plan** : entrez un vecteur de la taille du nombre de plans, contenant le nombre de combinaisons par plan souhaité.

Afficher les feuilles individuelles : activez cette option si vous voulez qu'une feuille pour chaque individu interrogé soit générée afin de l'utiliser à titre de feuille de réponse.

Affectation : si l'option précédente est activée, choisissez si vous voulez que les choix soient présentés toujours dans le même ordre (fixe) ou dans des ordres aléatoires pour les répondants (aléatoire).

Inclure des références : si l'option d'affichage des feuilles individuelles est activée, activez cette option si vous voulez que les feuilles du classeur soient liées en utilisant des formules Excel. Lorsqu'on entre une réponse dans la feuille individuelle, elle est automatiquement ajoutée au tableau général.

Résultats

Informations sur les variables : dans ce tableau sont récapitulées toutes les informations sur les attributs sélectionnés.

Plan d'analyse Max Diff : dans ce tableau sont affichés les choix générés. Les colonnes vides servent à être remplies avec les réponses (deux colonnes par individu). L'individu interrogé entrera le numéro du choix sélectionné (numéro associé à la colonne choisie). Si l'option « afficher les feuilles individuelles » est activée et que l'option « inclure des références » l'est aussi. Alors les colonnes vides du tableau font directement références aux feuilles associées à chaque individu.

Lancer l'analyse : une fois que tous les individus ont rempli le plan d'analyse MaxDiff, vous pouvez cliquer sur le bouton « Lancer l'analyse » afin d'ouvrir la boite de dialogue préremplie permettant d'effectuer l'analyse MaxDiff.

Feuilles _Res : ces feuilles sont les feuilles qui s'affichent lorsque l'option « afficher les feuilles individuelles » est activée. Chaque feuille correspond à un questionnaire avec le nom de l'analyse, le numéro du répondant ainsi qu'un tableau pour chaque choix (best and worst).

Exemple

Un exemple d'analyse MaxDiff basée sur le choix est disponible sur le Centre d'aide XLSTAT :

http://www.xlstat.com/demo-maxdifff.htm

Bibliographie

Louviere, J. J. (1991). Best-Worst Scaling: A Model for the Largest Difference Judgments, Working Paper, University of Alberta.

Marley, A.A.J. and Louviere, J.J. (2005). Some probabilistic models of best, worst, and best–worst choices. *Journal of Mathematical Psychology*, **49**, 464–480.

Analyse MaxDiff

Utilisez cet outil pour effectuer une analyse Max-Diff. Cet outil est inclus dans le module XLSTAT-Conjoint et doit être utilisé sur des plans d'expérience générés à l'aide de l'outil permettant de générer des plans pour l'analyse Max-Diff de XLSTAT-Conjoint.

Dans cette section :

Description

La méthode MaxDiff est une méthode introduite par Jordan Louvière qui permet de connaître l'importance d'attributs en utilisant une présentation originale. On présente au répondant une série d'attributs parmi lesquels celui-ci doit choisir le plus important et le moins important.

Cet outil permet de mener à bien l'étape consistant en l'analyse des résultats obtenus après le recueil des réponses obtenues auprès d'un échantillon de personnes.

L'analyse de ces choix peut se faire en utilisant un algorithme « logit » s'appuyant sur le modèle logit conditionnel qui permet d'obtenir une utilité pour les attributs, ou l'algorithme bayésien hiérarchique qui permet d'obtenir des résultats individu par individu.

Modèle logit conditionnel

La régression logistique conditionnelle est basée sur un modèle proche de celui de la régression logistique. La différence vient du fait que tous les individus sont soumis à différentes situations avant d'exprimer leur choix. Le fait de savoir que ce sont les mêmes individus qui ont répondu apporte de l'information que le modèle de régression logistique conditionnelle permet de prendre en compte (NB : les observations ne sont pas indépendantes à l'intérieur d'un même bloc correspondant au même individu).

La probabilité qu'un individu i choisisse le produit j est donnée par :

$$P_{ij} = \frac{e^{\beta^T z_{ij}}}{\sum_k e^{\beta^T z_{ik}}}$$

A partir de cette probabilité, on calcule une fonction de vraisemblance :

$$l(\beta) = \sum_{i=1}^{n} \sum_{j=1}^{J} y_{ij} \log(P_{ij})$$

Avec y variable binaire indiquant le choix de l'individu i pour le produit j et J nombre de choix offerts à chaque individu.

Pour estimer les paramètres β du modèle (les coefficients de la fonction linéaire), on cherche à maximiser la fonction de vraisemblance. Contrairement à la régression linéaire, une solution analytique exacte n'existe pas. Il est donc nécessaire d'utiliser un algorithme itératif. XLSTAT utilise un algorithme de Newton-Raphson.

Pour éviter la dépendance linéaire, on a défini arbitrairement l'utilité du premier élément à zéro et on estime les utilités des autres éléments par rapport à ce premier élément maintenu à zéro.

Algorithme Bayésien hiérarchique

Les différents paramètres sont estimés au niveau individuel via une procédure itérative (échantillonnage de Gibbs) qui tient compte du choix de chacun des individus ainsi que de la distribution globale de ces choix. Les estimations au niveau individuel permettent d'améliorer la précision des importances. L'analyse MaxDiff permet au travers du modèle Bayésien hiérarchique d'obtenir des importances pour chaque individu et chaque attribut.

Une fois le modèle Bayésien hiérarchique appliqué sur les résultats des questionnaires en multipliant par -1 les X pour le cas « pire », les coefficients du modèle sont transformés de manière à obtenir des scores MaxDiff.

Ceux-ci sont centrés puis transformés à l'aide de la formule $\frac{\exp(\beta)}{\exp(\beta)+nb_{alter}-1}$ avec nb_{alter} le nombre d'alternatives proposées lors de chaque choix proposé. Ce score est ensuite remis sur une échelle de façon à ce que la somme soit égale à 100.

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.

OK : cliquez sur ce bouton pour lancer les calculs.

Annuler : cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

Aide : cliquez sur ce bouton pour afficher l'aide.

 : cliquez sur ce bouton pour rétablir les options par défaut.

 : cliquez sur ce bouton pour effacer les sélections de données.

 : cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.

Onglet **Général** :

Chargement automatique des données  :

L'analyse MaxDiff nécessite le chargement de deux tableaux de données : un tableau contenant les réponses des individus et un tableau contenant les différents choix (ou combinaisons de choix). Si les réponses des individus sont rassemblées sur une feuille contenant un plan d'expérience pour l'analyse MaxDiff généré par XLSTAT, vous pouvez charger les deux tableaux de données automatiquement. Pour cela, il vous suffit de cliquer sur le bouton « baguette magique » puis de sélectionner n'importe quelle cellule de la feuille de résultat contenant le plan pour l'analyse MaxDiff généré par XLSTAT. Afin que les données soit chargées correctement il est important que vous n'ayez pas modifié manuellement la feuille de résultats contenant le plan généré par XLSTAT (pas d'ajouts de lignes ou de colonnes,…). Vous pouvez également charger les données en sélectionnant les différents tableaux séparément.

Réponses : sélectionnez les réponses des individus sous forme de nombres associés aux choix effectués. Ainsi s'il y a 3 possibilités, il faudra entrer soit 1, soit 2, soit 3. Si des en-têtes ont été sélectionnés, veuillez vérifier que l'option « Libellés des variables » est activée. Cette sélection correspond aux choix effectués par les individus, elle correspond à la partie droite du tableau « Plan d'analyse MaxDiff ».

Choix : sélectionnez les choix générés par l'outil de génération d'analyses MaxDiff d'XLSTAT-Conjoint. Si des en-têtes ont été sélectionnés, veuillez vérifier que l'option « Libellés des variables » est activée. La 1ère colonne comprenant les numéros des sélections ne doit pas être sélectionnée.

Terminologie : choisissez parmi les possibilités offertes celle qui correspond le mieux au cadre de votre analyse.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des variables : activez cette option si la première ligne des données sélectionnées contient un libellé.

Poids des réponses : activez cette option si vous voulez associer des poids aux réponses des individus. Sélectionnez les poids dans un vecteur vertical dont la taille est égale au nombre d'individus. Si des en-têtes ont été sélectionnés, veuillez vérifier que l'option « Libellés des variables » est activée.

Onglet **Options** :

Méthode : sélectionnez la méthode que vous désirez utiliser, parmi Logit ou Bayésien Hiérarchique.

Intervalle de confiance (%) : entrez l'étendue en pourcentage de l'intervalle de confiance à utiliser pour les différents tests et pour le calcul des intervalles de confiance autour des paramètres et des prédictions. Valeur par défaut : 95%.

Options du bayésien : le nombre d'itérations pour la période de chauffe et le temps maximal de l'algorithme peuvent être réglés.

Conditions d'arrêt : le nombre d'itérations et l'indice de convergence pour l'algorithme peuvent être réglés. Si le nombre d'itérations est atteint et si $Abs(moyenne(BetaOld - BetaNew))$ (où $BetaOld$ défini la valeur des coefficients à l'itération $k-1$ et $BetaNew$ défini leur valeur à l'itération k) n'atteint pas la valeur de convergence alors l'algorithme est arrêté.

Onglet **Données manquantes** :

Ne pas accepter les valeurs manquantes : activez cette option pour que XLSTAT empêche la poursuite des calculs si des valeurs manquantes sont détectées.

Supprimer les observations : activez cette option pour supprimer les observations comportant des données manquantes.

Résultats

Informations sur les variables : dans ce tableau toutes les informations sur les attributs sélectionnés sont récapitulées.

Analyse des comptages : ces tableaux sont présentés pour l'ensemble des individus puis pour chaque individu. Ils présentent le nombre de fois où chaque produit a été choisi comme meilleur, comme pire, et la différence des deux.

Utilités : si le modèle choisi est le « logit », ce tableau est affiché pour permettre d'évaluer l'utilité associée à chaque attribut.

Les résultats suivants sont affichés dans le cas du modèle Bayésien Hiérarchique.

Scores MaxDiff : dans ce tableau sont affichés les scores associés à chaque attribut et pour chaque individu. Des tableaux de statistiques descriptives pour tous les individus sont aussi disponibles.

Coefficients du modèle : dans ce tableau les coefficients du modèle HB sont donnés de manière brute.

Coefficients d'ajustement : dans ce tableau est affichée une série de statistiques pour le modèle indépendant (correspondant au cas où la combinaison linéaire des variables explicatives se réduit à une constante) et pour le modèle ajusté.

- **Observations** : le nombre total d'observations prises en compte (somme des poids des observations);
- **Somme des poids** : le nombre total d'observations prises en compte (somme des poids des observations multipliés par les poids dans la régression);
- **-2 Log(Vrais.)** : le logarithme de la fonction de vraisemblance associée au modèle;
- **rlh** : racine de la vraisemblance. Cette valeur varie entre 0 et 1, 1 correspond à un modèle qui ajuste parfaitement les données.
- **rlh par individu** : La valeur RLH (Root LikeliHood) est un indice de 0 à 1. Plus la valeur RLH d'un répondant est élevée, plus le répondant a répondu de manière cohérente aux questions de choix.

Les résultats sont ensuite donnés pour chaque individu.

Exemple

Un exemple d'analyse Max-Diff est disponible sur le Centre d'aide XLSTAT :

http://www.xlstat.com/demo-maxdifff.htm

Bibliographie

Louviere, J. J. (1991). Best-Worst Scaling: A Model for the Largest Difference Judgments, Working Paper, University of Alberta.

Marley, A.A.J. and Louviere, J.J. (2005). Some probabilistic models of best, worst, and best–worst choices. *Journal of Mathematical Psychology*, **49**, 464–480.

MONANOVA (Régression monotone)

Utilisez cet outil pour appliquer une régression monotone ou la méthode MONANOVA. Des options avancées vous permettent de choisir les contraintes sur le modèle et de tenir compte des interactions entre les facteurs. Cet outil est inclus dans le module XLSTAT-Conjoint.

Dans cette section :

Description

La régression monotone et la méthode MONANOVA diffèrent uniquement dans le fait que les variables explicatives sont soit quantitatives, soit qualitatives. Ces méthodes sont basées sur des algorithmes itératifs issus de l'algorithme ALS (moindres carrés alternés). Leur principe est simple, il consiste en l'alternance entre une estimation classique du type régression linéaire ou ANOVA et d'une transformation monotone des variables dépendantes (issue des recherches sur l'optimal scaling).

L'algorithme MONANOVA a été présenté par Kruskal (1965). Les méthodes de régression monotone et les travaux sur l'algorithme ALS sont quant à eux dus à Young et al. (1976).

Ces méthodes sont utilisées couramment dans le cadre de l'analyse conjointe basée sur les profils complets. XLSTAT-Conjoint permet de les appliquer à l'intérieur d'une analyse conjointe (voir le chapitre sur l'analyse conjointe basée sur les profils complets) mais aussi de manière indépendante.

L'outil régression monotone (MONANOVA) permet de combiner une transformation monotone des réponses à une régression linéaire de manière à améliorer les résultats de la régression.

XLSTAT-Conjoint permet d'ajouter des interactions et de faire varier les contraintes sur les variables qualitatives.

Méthode :

La régression monotone combine une étape de régression linéaire ordinaire entre les variables explicatives et la variable réponse et une étape de transformation de la variable réponse de manière à optimiser la qualité de prédiction.

L'algorithme du type moindres carrés alternés est le suivant :

1. Régression OLS entre la variable réponse Y et les variables explicatives X. On obtient les coefficients β.

2. Calcul des valeurs prédites de Y avec le modèle obtenu $Pred(Y) = \beta * X$

3. Transformation de Y en utilisant une transformation monotone (Kruskal, 1965) de façon à ce que $Pred(Y)$ et Y soient proches (utilisation de l'optimal scaling).

4. Régression OLS entre Y_trans et les variables explicatives X. On obtient de nouvelles valeurs pour les β.

5. Les étapes 2 à 4 sont répétées jusqu'à ce que la variation du R^2 d'une étape à l'autre soit plus petite que le critère de convergence.

Coefficients d'ajustement (MONANOVA)

Dans le cadre d'une régression monotone, des résultats supplémentaires sont disponibles. Ces résultats sont généralement associés à une analyse multivariée mais comme nous sommes dans le cas d'une transformation leur présence est aussi nécessaire. Au lieu d'utiliser le carré des corrélations canoniques entre les mesures, nous utilisons le R^2 entre la mesure et sa transformation car il y a une seule transformation linéaire.

Ainsi, XLSTAT-Conjoint calcule le lambda de Wilks, la trace de Pillai, le trace de Hotelling-Lawlet et la plus grande racine de Roy en utilisant une matrice avec comme plus grande valeur propre le R^2 et 0 pour les autres. La plus grande racine de Roy donne une borne inférieure pour la p-valeur. Les autres indices donnent des bornes supérieures à la p-valeur du modèle.

Interactions

On désigne par interaction un facteur artificiel (non mesuré) reflétant l'interaction entre au moins deux facteurs mesurés. Par exemple, si on applique un traitement à une plante, et que les essais sont réalisés sous deux intensités lumineuses différentes, on pourra inclure dans le modèle un facteur d'interaction traitement*lumière qui permettra d'identifier une éventuelle interaction entre les deux facteurs. S'il y a une interaction entre les deux facteurs, on observera sur les plantes un effet significativement plus important lorsque la lumière est forte et que le traitement est de type 2, alors que l'effet est moyen pour les couples (lumière faible, traitement 2) et (lumière forte, traitement 1).

Pour faire un parallèle avec la régression linéaire, les interactions sont équivalentes à des produits entre les valeurs explicatives continues, bien qu'ici l'obtention des interactions nécessite plus qu'une simple multiplication entre deux variables. Néanmoins la notation utilisée pour représenter l'interaction entre le facteur A et le facteur B est $A * B$.

XLSTAT permet de facilement définir les interactions à prendre en compte dans le modèle.

Contraintes

Lorsque des variables qualitatives sont utilisées (MONANOVA), on les nomme alors facteurs. Au cours des calculs, chaque facteur est décomposé en une sous-matrice contenant autant de colonnes qu'il y a de modalités dans le facteur. Typiquement, il s'agit d'un tableau disjonctif complet. Cette décomposition pose néanmoins un problème : s'il y a g modalités, le rang de cette sous-matrice n'est pas g mais $g-1$. Cela entraîne la nécessité de supprimer l'une des colonnes de la sous-matrice, et éventuellement de transformer les autres colonnes. Plusieurs stratégies sont possibles en fonction de l'interprétation que l'on veut ensuite faire:

1) **a1=0** : le paramètre correspondant à la première modalité est nul. Ce choix permet d'imposer que l'effet de la première modalité correspond à un standard. Dans ce cas, la constante du modèle est égale à la moyenne de la variable dépendante pour le groupe 1.

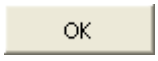
2) **an=0** : le paramètre correspondant à la dernière modalité est nul. Ce choix permet d'imposer que l'effet de la dernière modalité correspond à un standard. Dans ce cas, la constante du modèle est égale à la moyenne de la variable dépendante pour le groupe g.

3) **Somme(ai)=0** : la somme des paramètres est nulle. Ce choix permet d'imposer que la constante du modèle est égale à la moyenne de la variable dépendante lorsque le modèle est équilibré.

Remarque : si le choix de la contrainte influence la valeur des paramètres, il n'en a aucun sur les valeurs prédites et sur les différentes statistiques d'ajustement.

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.

Onglet **Général** :

Y / Variables dépendantes :

Quantitatives : sélectionnez la ou les variables réponse que vous souhaitez modéliser. Si plusieurs variables sont sélectionnées, XLSTAT fera les calculs pour chacune des variables indépendamment. Si des en-têtes de colonnes ont été sélectionnés, veuillez vérifier que l'option « Libellés des variables » est activée.

X / Variables explicatives :

Quantitatives : activez cette option si vous voulez inclure une ou plusieurs variables explicatives quantitatives dans le modèle. Sélectionnez alors la ou les variables correspondantes sur la feuille Excel. Les données sélectionnées doivent être de type numérique. Si le libellé des variables a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

Qualitatives : sélectionnez la ou les variables explicatives qualitatives (les facteurs) sur la feuille Excel. Les données sélectionnées peuvent être de tout type, mais les données numériques sont automatiquement considérées comme nominales. Si le libellé des variables a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des variables : activez cette option si la première ligne des données sélectionnées (variables dépendantes et explicatives, libellés des observations, poids) contient un libellé.

Libellés des observations : activez cette option si vous voulez utiliser des libellés d'observations pour l'affichage des résultats. Si l'option « Libellés des variables » est activée, la première cellule de la sélection doit comprendre un en-tête. Si vous n'activez pas cette option, des libellés seront automatiquement créés (Obs1, Obs2,…).

Poids des observations : activez cette option si vous voulez pondérer les observations. Si vous n'activez pas cette option, les poids seront tous considérés comme valant 1. Les poids doivent être impérativement supérieurs ou égaux à 0. Un poids de 2 est équivalent à répéter deux fois la même observation. Si un en-tête de colonne a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

Onglet **Options**:

Constante fixée : activez cette option pour fixer la constante du modèle de régression à une valeur que vous devez ensuite saisir (0 par défaut).

Tolérance : activez cette option pour permettre à l'algorithme de calcul de la régression OLS ne pas prendre en compte les variables qui seraient soit constantes soit trop corrélées avec d'autres variables déjà utilisées dans le modèle (0.0001 par défaut).

Interactions / Niveau : activez cette option pour inclure des interactions dans le modèle puis entrez le niveau maximum d'interaction (valeur comprise entre 1 et 4).

Intervalle de confiance (%) : entrez l'étendue en pourcentage de l'intervalle de confiance à utiliser pour les différents tests, et pour le calcul des intervalles de confiance autour des paramètres et des prédictions. Valeur par défaut : 95%.

Contraintes : des détails sur les différentes options sont disponibles dans la section description.

a1 = 0 : choisissez cette option pour que le paramètre de la première modalité de chaque facteur soit fixé à 0.

an = 0 : choisissez cette option pour que le paramètre de la dernière modalité de chaque facteur soit fixé à 0.

Somme (ai) = 0 : pour chaque facteur la somme des paramètres associés aux différentes modalités vaut 0.

Conditions d'arrêt : le nombre d'itérations et l'indice de convergence pour l'algorithme MONANOVA peuvent être réglés.

Onglet **Données manquantes** :

Supprimer les observations : activez cette option pour supprimer les observations comportant des données manquantes.

Estimer les données manquantes : activez cette option pour estimer les données manquantes avant le début des calculs.

- **Moyenne ou mode** : activez cette option pour estimer les données manquantes en utilisant la moyenne (variables quantitatives) ou le mode (variables qualitatives) pour les variables correspondantes.
- **Plus proche voisin** : activez cette option pour estimer les données manquantes d'une observation en recherchant le plus proche voisin de l'observation.

Onglet **Sorties** :

Statistiques descriptives : activez cette option pour afficher les statistiques descriptives pour les variables sélectionnées.

Corrélations : activez cette option pour afficher la matrice de corrélation pour les variables quantitatives (dépendantes et explicatives).

Analyse de la variance : activez cette option pour afficher le tableau d'analyse de la variance.

Type I/II/III SS : activez cette option pour afficher les tableaux de l'analyse de la variance de Type I, II et III (*Type I/II/III Sum of Squares*).

Coefficients normalisés : activez cette option pour afficher les paramètres normalisés du modèle (coefficients bêta).

Prédictions et résidus : activez cette option pour afficher les prédictions et les résidus pour l'ensemble des observations.

Onglet **Graphiques** :

Graphiques de régression : activez cette option pour afficher les graphiques de régression :

- **Coefficients normalisés** : activez cette option pour afficher sur un graphique les paramètres normalisés du modèle avec leur intervalle de confiance.

Graphiques de transformation : activez cette option pour afficher les graphiques de transformation monotone des données.

Résultats

Statistiques descriptives : les tableaux de statistiques descriptives présentent pour toutes les variables sélectionnées des statistiques simples. Pour les variables dépendantes (en bleu) et les variables explicatives quantitatives, sont affichés le nombre d'observations, le nombre de données manquantes, le nombre de données non manquantes, la moyenne, et l'écart-type (non biaisé). Pour les variables explicatives qualitatives sont affichés le nom des différentes modalités ainsi que leur fréquence respective.

Matrice de corrélation : ce tableau est affiché afin de vous permettre d'avoir un aperçu des corrélations entre les différentes variables sélectionnées.

Coefficients d'ajustement : dans ce tableau sont affichées les statistiques relatives à l'ajustement du modèle de régression :

- **Observations** : le nombre d'observations prises en compte dans les calculs. Dans les formules présentées ci-dessous n désigne le nombre d'observations.
- **Somme des poids** : la somme des poids des observations prises en compte dans les calculs. Dans les formules présentées ci-dessous W désigne la somme des poids.
- **DDL** : le nombre de degrés de liberté pour le modèle retenu (correspondant à la partie erreurs).
- R^2 : le coefficient de détermination du modèle. Ce coefficient, dont la valeur est comprise entre 0 et 1, n'est affiché que si la constante du modèle n'est pas fixée par l'utilisateur. Sa valeur est définie par

$$R^2 = 1 - \frac{\sum_{i=1}^{n} w_i (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} w_i (y_i - \bar{y}_i)^2}, \text{ avec } \bar{y} = \frac{1}{n} \sum_{i=1}^{n} w_i y_i$$

- Le R^2 s'interprète comme la proportion de la variabilité de la variable dépendante expliquée par le modèle. Plus le R^2 est proche de 1, meilleur est le modèle. L'inconvénient du R^2 est qu'il ne prend pas en compte le nombre de variables utilisées pour ajuster le modèle.

- R^2 **ajusté** : le coefficient de détermination ajusté du modèle. Le R^2 ajusté peut être négatif si le R^2 est voisin de zéro. Ce coefficient n'est affiché que si la constante du modèle n'est pas fixée par l'utilisateur. Sa valeur est définie par

$$\hat{R}^2 = 1 - (1 - R^2) \frac{W - 1}{W - p - 1}$$

Le R^2 ajusté est une correction du R^2 qui permet de prendre en compte le nombre de variables utilisées dans le modèle.

- **MCE** : la moyenne des carrés des erreurs (MCE) est définie par :

$$MCE = \frac{1}{W - p^*} \sum_{i=1}^{n} w_i (y_i - \hat{y}_i)^2$$

- **RMCE** : la racine de la moyenne des carrés des erreurs (RMCE) est la racine carrée de la MCE.

- **MAPE** : la *Mean Absolute Percentage Error* est calculée comme suit :

$$MAPE = \frac{100}{W} \sum_{i=1}^{n} w_i \left| \frac{y_i - \hat{y}_i}{y_i} \right|$$

- **DW** : le coefficient de Durbin-Watson est défini par

$$DW = \frac{\sum_{i=2}^{n} [(y_i - \hat{y}_i) - (y_{i-1} - \hat{y}_{i-1})]^2}{\sum_{i=1}^{n} w_i (y_i - \hat{y}_i)^2}$$

Ce coefficient correspond au coefficient d'autocorrélation d'ordre 1 et permet de vérifier que les résidus du modèle ne sont pas autocorrélés, sachant que l'indépendance des résidus est l'une des hypothèses de base de la régression linéaire. L'utilisateur pourra se référer à une table des coefficients de Durbin-Watson pour vérifier si l'hypothèse d'indépendance des résidus est acceptable.

- Cp : le coefficient Cp de Mallows est défini par

$$Cp = \frac{SCE}{\hat{\sigma}} + 2p^* - W$$

où SCE est la somme du carré des erreurs pour le modèle avec p variables explicatives, et où $\hat{\sigma}$ correspond à l'estimateur de la variance des résidus pour le modèle comprenant toutes les

variables explicatives. Plus le coefficient Cp est proche de $p*$ moins le modèle est biaisé.

- **AIC** : le critère d'information d'Akaike (Akaike's Information Criterion) est défini par

$$AIC = W \ln\left(\frac{SCE}{W}\right) + 2p^*$$

Ce critère proposé par Akaike (1973) dérive de la théorie de l'information, et s'appuie sur la mesure de Kullback et Leibler (1951). C'est un critère de sélection de modèles qui pénalise les modèles pour lesquels l'ajout de nouvelles variables explicatives n'apporte pas suffisamment d'information au modèle, l'information étant mesurée au travers de la SCE. On cherche à minimiser le critère AIC.

- **SBC** : le critère bayésien de Schwarz (Schwarz's Bayesian Criterion) est défini par

$$SBC = W \ln\left(\frac{SCE}{W}\right) + \ln(W)p^*$$

Ce critère proposé par Schwarz (1978) est proche du critère AIC, et comme ce dernier on cherche à le minimiser.

- **PC** : le critère de prédiction d'Amemiya (Amemiya's Prediction Criterion) est défini par

$$PC = \frac{(1-R^2)(W+p^*)}{W-p^*}$$

Ce critère proposé par Amemiya (1980) permet comme le R^2 ajusté de tenir compte de la parcimonie du modèle.

- **Press** : la statistique du Press (predicted residual error sum of squares) n'est affichée que si l'option correspondante a été activée dans la boîte de dialogue. Elle est définie par

$$Press = \sum_{i=1}^{n} w_i (y_i - \hat{y}_{i(-i)})^2$$

où $\hat{y}_{i(-i)}$ est la prédiction pour l'observation i lorsque cette dernière n'est pas utilisée pour l'estimation des paramètres. On obtient alors

$$\text{Press RMCE } = \sqrt{\frac{\text{Press}}{W-p^*}}$$

Le Press RMCE peut alors être comparé au RMCE. Une différence importante entre les deux indique que le modèle est sensible à la présence ou absence de certaines observations dans le modèle.

- **Itérations** : nombre d'itérations nécessaires à la convergence.

Coefficients d'ajustement (MONANOVA) : dans ce tableau sont affichées les statistiques relatives à l'ajustement du modèle de régression spécifique au cas de l'ANOVA monotone. Ces indices sont le lambda de Wilks, la trace de Pillai, le trace de Hotelling-Lawlet et la plus grande

racine de Roy. Pour plus de détails sur ces statistiques, on peut voir la partie description de cette aide.

Le **tableau d'analyse de la variance** permet d'évaluer le pouvoir explicatif des variables explicatives. Dans le cas où la constante du modèle n'est pas fixée à une valeur donnée, le pouvoir explicatif est évalué en comparant l'ajustement (au sens des moindres carrés) du modèle final avec l'ajustement du modèle rudimentaire composé d'une constante égale à la moyenne de la variable dépendante. Dans le cas où la constante du modèle est fixée, la comparaison est faite par rapport au modèle pour lequel la variable dépendante serait égale à la constante fixée.

Les tableaux des **Type I/II/III SS** permettent de visualiser l'influence du retrait d'une variable explicative sur l'ajustement du modèle, toutes les autres variables étant conservées, au sens de la somme des carrés des erreurs (SCE), de la moyenne des carrés des erreurs, du F de Fisher, ou de la probabilité associée au F de Fisher. Plus la probabilité est faible, plus la contribution de la variable au modèle est importante, toutes les autres variables étant déjà dans le modèle.

Le tableau des **coefficients normalisés** (aussi appelés coefficients β) permet de comparer le poids relatif des variables. Plus la valeur absolue d'un coefficient est élevée, plus le poids de la variable correspondante est important. Lorsque l'intervalle de confiance autour des coefficients normalisés comprend la valeur 0 (cela est facilement visible sur le graphique des coefficients normalisés), le poids d'une variable dans le modèle n'est pas significatif.

Dans le tableau des **prédictions et résidus** sont donnés pour chaque observation, son poids, la valeur observée de la variable dépendante, la valeur de la variable dépendante transformée, la prédiction du modèle, les résidus et les intervalles de confiance. Deux types d'intervalles de confiance sont affichés : un intervalle de confiance autour de la moyenne (correspondant au cas où l'on ferait la prédiction pour un nombre infini d'observations avec un ensemble de valeurs données des variables explicatives) et un intervalle autour de la prédiction ponctuelle (correspondant au cas d'une prédiction isolée pour des valeurs données des variables explicatives). Le second intervalle est toujours plus grand que le premier, les aléas étant plus importants.

Le **graphique** qui suit permet de visualiser les transformations monotones obtenues par l'algorithme MONANOVA.

Exemple

Un exemple de MONANOVA est disponible sur le Centre d'aide XLSTAT :

http://www.xlstat.com/demo-monanovaf.htm

Bibliographie

Kruskal, J. B. (1965). Analysis of Factorial Experiments by Estimating Monotone Transformations of the Data. *Journal of the Royal Statistical Society. Series B (Methodological).* **27(2)**, 251-263.

Sahai H. and Ageel M.I. (2000). The Analysis of Variance. Birkhaüser, Boston.

Takane Y., Young F. W. and De Leeuw J. (1977). Nonmetric individual differences multidimensional scaling: an alternating least squares method with optimal scaling features. *Psychometrika*, **42**, 7-67.

Young F. W., De Leeuw J. and Takane Y. (1976). Regression with qualitative and quantitative variables: alternating least squares method with optimal scaling features. *Psychometrika*, **41**, 505-529.

Modèle logit conditionnel (régression logistique conditionnelle)

Utilisez cet outil pour appliquer une régression logistique conditionnelle (modèle logit conditionnel). Des options avancées vous permettent de choisir les contraintes sur le modèle et de tenir compte des interactions entre les facteurs. Cet outil est inclus dans le module XLSTAT-Conjoint.

Dans cette section :

Description

La régression logistique conditionnelle fait partie du module XLSTAT-Conjoint.

La régression logistique conditionnelle est une méthode surtout utilisée dans sa forme évoluée dans le cadre de l'analyse conjointe, elle est néanmoins utile lorsqu'on analyse un certain type de données. C'est McFadden (1973) qui a introduit ce modèle. Au lieu d'avoir une ligne par individu, on aura une ligne par choix possible. Ainsi, ce ne sont plus les caractéristiques des individus qui sont modélisées mais celles des différentes alternatives. Ainsi, si on cherche à étudier des habitudes de transport, par exemple, on aura quatre types de transports (voiture / train / avion / vélo), chacun de ces type de transport à des caractéristiques (son prix, son coût environnemental…), mais un individu ne choisira qu'un seul des quatre moyens de transport. Dans le cadre d'un modèle logit conditionnel, on aura pour N individus, $N * 4$ lignes avec 4 lignes pour chaque individu associé à chacun des moyens de transport. La variable réponse binaire indiquera le choix de l'individu (1) et 0 si l'individu n'a pas choisi cette option. Il faudra aussi sélectionner une colonne associée au nom des individus (avec 4 lignes par individu pour l'exemple des moyens de transport). Les variables explicatives devront aussi avoir $N * 4$ lignes.

Le modèle

La régression logistique conditionnelle est basée sur un modèle proche de celui de la régression logistique. La différence vient de ce que tous les individus sont soumis à différentes situations avant d'exprimer leur choix (sous forme d'une variable binaire qui constitue la variable dépendante). Le fait de savoir que ce sont les mêmes individus qui ont répondu apporte de l'information que le modèle de régression logistique conditionnelle permet de prendre en

compte (NB : les observations ne sont pas indépendantes à l'intérieur d'un même bloc correspondant au même individu).

La probabilité qu'un individu i choisisse le produit j est donnée par :

$$P_{ij} = \frac{e^{\beta^T z_{ij}}}{\sum_k e^{\beta^T z_{ik}}}$$

A partir de cette probabilité, on calcule une fonction de vraisemblance :

$$l(\beta) = \sum_{i=1}^{n} \sum_{j=1}^{J} y_{ij} \log(P_{ij})$$

Avec y variable binaire indiquant le choix de l'individu i pour le produit j et J nombre de choix offerts à chaque individu.

Pour estimer les paramètres β du modèle (les coefficients de la fonction linéaire), on cherche à maximiser la fonction de vraisemblance. Contrairement à la régression linéaire, une solution analytique exacte n'existe pas. Il est donc nécessaire d'utiliser un algorithme itératif. XLSTAT utilise un algorithme de Newton-Raphson.

Coefficients d'ajustement (logit conditionnel)

Un certain nombre de coefficients d'ajustement spécifiques sont aussi obtenus :

- Rapport de vraisemblance R : $R = -2(\log(L) - \log(L_0))$

- Borne supérieure du rapport de vraisemblance U : $U = -2\log(L_0)$

- Aldrich-Nelson : $AN = \frac{R}{R+N}$

- Cragg-Uhler 1 : $CU_1 = 1 - e^{-\frac{R}{N}}$

- Cragg-Uhler 2 : $CU_2 = \frac{1-e^{-\frac{R}{N}}}{1-e^{-\frac{U}{N}}}$

- Estrella : $Estrella = 1 - (1 - \frac{R}{U})^{\frac{U}{N}}$

- Estrella ajusté : $Adj.Estrella = 1 - \left(\frac{\log(L)-k}{\log(L_0)}\right)^{\frac{2}{N}\log(L_0)}$

- Veall-Zimmermann : $VZ = \frac{R(U+N)}{U(R+N)}$

Avec N taille de l'échantillon total et K nombre de variables explicatives.

Contraintes pour les variables qualitatives

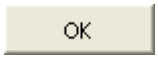
Au cours des calculs, chaque facteur est décomposé en une sous-matrice contenant autant de colonnes qu'il y a de modalités dans le facteur. Typiquement, il s'agit d'un tableau disjonctif complet. Cette décomposition pose néanmoins un problème : s'il y a g modalités, le rang de cette sous- matrice n'est pas g mais $g-1$. Cela entraîne la nécessité de supprimer l'une des colonnes de la sous-matrice, et éventuellement de transformer les autres colonnes. Plusieurs stratégies sont possibles en fonction de l'interprétation que l'on veut ensuite faire :

1) **a1=0** : le paramètre correspondant à la première modalité est nul. Ce choix permet d'imposer que l'effet de la première modalité correspond à un standard.

2) **Somme(ai)=0** : la somme des paramètres est nulle.

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.

Onglet **Général** :

Variable réponse : sélectionnez la variable réponse que vous souhaitez modéliser. Si des en-têtes de colonnes ont été sélectionnés, veuillez vérifier que l'option « Libellés des variables » est activée.

Variable sujets : sélectionnez la variable sujets qui comprend le nom des individus. Si des en-têtes de colonnes ont été sélectionnés, veuillez vérifier que l'option « Libellés des variables » est activée.

Quantitatives : activez cette option si vous voulez inclure une ou plusieurs variables explicatives quantitatives dans le modèle. Sélectionnez alors la ou les variables correspondantes sur la feuille Excel. Les données sélectionnées doivent être de type numérique. Si le libellé des variables a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

Qualitatives : sélectionnez la ou les variables explicatives qualitatives sur la feuille Excel. Les données sélectionnées peuvent être de tout type, mais les données numériques sont automatiquement considérées comme nominales. Si le libellé des variables a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des variables : activez cette option si la première ligne des données sélectionnées (variables dépendantes et explicatives, libellés des observations, poids) contient un libellé.

Poids des observations : activez cette option si vous voulez pondérer les observations. Si vous n'activez pas cette option, les poids seront tous considérés comme valant 1. Les poids doivent être impérativement supérieurs ou égaux à 0. Un poids de 2 est équivalent à répéter deux fois la même observation. Si un en-tête de colonne a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

Onglet **Options** :

Tolérance : entrez la valeur de la tolérance seuil en deçà de laquelle une variable est automatiquement ignorée.

Interactions / Niveau : activez cette option pour inclure des interactions dans le modèle puis entrez le niveau maximum d'interaction (valeur comprise entre 1 et 4).

Intervalle de confiance (%) : entrez l'étendue en pourcentage de l'intervalle de confiance à utiliser pour les différents tests, et pour le calcul des intervalles de confiance autour des paramètres et des prédictions. Valeur par défaut : 95%.

Contraintes : des détails sur les différentes options sont disponibles dans la section description.

a1 = 0 : choisissez cette option pour que le paramètre de la première modalité de chaque facteur soit fixé à 0.

Somme (ai) = 0 : pour chaque facteur la somme des paramètres associés aux différentes modalités vaut 0.

Conditions d'arrêt :

- **Itérations** : entrez le nombre maximal d'itérations pour l'algorithme de Newton-Raphson. Les calculs sont interrompus dès que le nombre maximal d'itérations est dépassé. Valeur par défaut : 100.
- **Convergence** : entrez la valeur seuil d'évolution maximale de log vraisemblance d'une itération à l'autre, qui une fois atteinte permet de considérer que l'algorithme a convergé. Valeur par défaut : 0,000001.

Onglet **Données manquantes** :

Supprimer les observations : activez cette option pour supprimer les observations comportant des données manquantes.

Estimer les données manquantes : activez cette option pour estimer les données manquantes avant le début des calculs.

- **Moyenne ou mode** : activez cette option pour estimer les données manquantes en utilisant la moyenne (variables quantitatives) ou le mode (variables qualitatives) pour les variables correspondantes.
- **Plus proche voisin** : activez cette option pour estimer les données manquantes d'une observation en recherchant le plus proche voisin de l'observation.

Onglet **Sorties** :

Statistiques descriptives : activez cette option pour afficher les statistiques descriptives pour les variables sélectionnées.

Corrélations : activez cette option pour afficher la matrice de corrélations des variables explicatives.

Coefficients d'ajustement : activez cette option pour afficher le tableau des statistiques d'ajustement du modèle.

Analyse de type III : activez cette option pour afficher le tableau d'analyse de la variable de type III.

Coefficients normalisés : activez cette option pour afficher les paramètres normalisés du modèle (coefficients bêta).

Prédictions et résidus : activez cette option pour afficher les prédictions et les résidus pour l'ensemble des observations.

Onglet **Graphiques** :

Graphiques de régression : activez cette option pour afficher les graphiques de régression :

- **Coefficients normalisés** : activez cette option pour afficher sur un graphique les paramètres normalisés du modèle avec leur intervalle de confiance.

Résultats

XLSTAT propose un nombre important de tableaux et de graphiques afin de faciliter l'analyse et l'interprétation des résultats.

Statistiques descriptives : les tableaux de statistiques descriptives présentent pour toutes les variables sélectionnées des statistiques simples. Pour les variables dépendantes (en bleu) et les variables explicatives quantitatives, sont affichés le nombre d'observations, le nombre de données manquantes, le nombre de données non manquantes, la moyenne, et l'écart-type (non biaisé). Pour les variables explicatives qualitatives sont affichés le nom des différentes modalités ainsi que leur fréquence respective.

Matrice de corrélation : ce tableau est affiché afin de vous permettre d'avoir un aperçu des corrélations entre les différentes variables sélectionnées.

Coefficients d'ajustement : dans ce tableau est affichée une série de statistiques pour le modèle indépendant (correspondant au cas où la combinaison linéaire des variables explicatives se réduit à une constante) et pour le modèle ajusté.

- **Observations** : le nombre total d'observations prises en compte (somme des poids des observations);
- **Somme des poids** : le nombre total d'observations prises en compte (somme des poids des observations multipliés par les poids dans la régression);
- **DDL** : degrés de liberté;
- **-2 Log(Vrais.)** : le logarithme de la fonction de vraisemblance associée au modèle;
- R^2**(McFadden)** : coefficient compris comme le R^2 entre 0 et 1 qui mesure le bon ajustement du modèle. Ce coefficient est égal à 1 moins le rapport de la vraisemblance du modèle ajusté sur la vraisemblance du modèle indépendant;
- R^2**(Cox et Snell) :** coefficient compris comme le R^2 entre 0 et 1 qui mesure le bon ajustement du modèle. Ce coefficient est égal à 1 moins le rapport de la vraisemblance du modèle ajusté sur la vraisemblance du modèle indépendant, le rapport étant porté à l'exposant $\frac{2}{S_w}$, où S_w est la somme des poids;
- R^2**(Nagelkerke) :** coefficient compris comme le R^2 entre 0 et 1 qui mesure le bon ajustement du modèle. Ce coefficient est égal au rapport du R^2 de Cox et Snell, divisé par 1 moins le la vraisemblance du modèle indépendant portée à l'exposant $\frac{2}{S_w}$;
- **AIC** : le critère d'information d'Akaike (Akaike's Information Criterion);
- **SBC** : le critère bayésien de Schwarz (Schwarz's Bayesian Criterion).
- **Itérations** : nombre d'itérations nécessaires à la convergence.

Coefficients d'ajustement (Logit conditionnel) : dans ce tableau sont affichées les statistiques relatives à l'ajustement du modèle de régression spécifique au cas du modèle logit conditionnel. Pour plus de détails sur ces statistiques, on peut voir la partie description de cette aide.

Test de l'hypothèse nulle H_0 : $Y = p_0$: l'hypothèse H_0 correspond au modèle indépendant qui donne la probabilité p_0 quelques soient les valeurs des variables explicatives ; on cherche à vérifier si le modèle ajusté est significativement plus performant que ce modèle. Trois tests sont proposés : le test du rapport des vraisemblance (-2 Log(Vrais.)), le test du Score, et le test test de Wald. Les trois statistiques suivent une loi du χ^2 dont les degrés de liberté sont indiqués.

Analyse de Type III : ce tableau n'a d'intérêt que s'il y a plus d'une variable explicative. On test ici le modèle ajusté contre un test dont on aurait retiré la variable de la ligne du tableau en question. Si la probabilité $Pr > LR$ est inférieur à un seul de signification que l'on se fixe (typiquement 0.05), alors la contribution de la variable à l'ajustement du modèle est significative. Sinon, elle peut être retirée du modèle.

Le tableau des **coefficients normalisés** (aussi appelés coefficients bêta) permet de comparer le poids relatif des variables. Plus la valeur absolue d'un coefficient est élevée, plus le poids de la variable correspondante est important. Lorsque l'intervalle de confiance autour des coefficients normalisés comprend la valeur 0 (cela est facilement visible sur le graphique des coefficients normalisés), le poids d'une variable dans le modèle n'est pas significatif.

Dans le tableau des **prédictions et résidus** sont donnés pour chaque observation, son poids, la valeur observée de la variable dépendante, la prédiction du modèle et la probabilité associé au modèle indépendant. Les mêmes valeurs divisées par le poids et les résidus standardisés.

Exemple

Un exemple de régression logistique conditionnelle est disponible sur le Centre d'aide XLSTAT :

http://www.xlstat.com/demo-clogitf.htm

Bibliographie

Ben-Akiva M. and Lerman S.R. (1985). Discrete Choice Analysis, The MIT Press.

McFadden D. (1974). Conditional logit analysis of qualitative choice behavior. In P. Zarembka (Ed.), Frontiers in Econometrics, Academic Press, 105-142.

Text mining

Extraction de caractéristique

L'extraction de caractéristique est utilisée pour réduire la quantité de ressources requises pour décrire un grand nombre de données textuelles. Il s'agit d'un terme générique qui décrit les méthodes de construction de combinaisons de variables destinées à résoudre cette problématique tout en décrivant les données avec une précision suffisante. Les "caractéristiques extraites" sont couramment utilisées dans les méthodes de classification de documents dans lesquelles la fréquence d'occurrence de chaque mot dans un document est utilisée comme caractéristique pour l'apprentissage d'un classifieur.

Dans cette section :

Description

Puisque les algorithmes d'apprentissage automatique fonctionnent selon un espace de caractéristiques numériques, l'entrée attendue est un tableau à deux dimensions où les lignes sont des observations (documents) à classer et les colonnes sont des caractéristiques (mots). Par conséquent, pour effectuer un apprentissage automatique sur des données textuelles, nous devons transformer nos documents selon une représentation vectorielle telle que nous puissions appliquer un apprentissage automatique numérique. Le processus d'encodage de documents dans cet espace de caractéristiques numériques est appelé *extraction de caractéristique* ou plus simplement vectorisation et constitue une première étape essentielle dans le traitement automatique du langage naturel (TALN).

Une représentation simple issue du processus de vectorisation est le modèle du « sac de mots », aussi connu sous le nom de modèle sémantique vectoriel. Avec ce schéma, un texte (comme une phrase ou un document) est représenté par l'ensemble formé de ses mots, sans tenir compte de la grammaire ni même de l'ordre des mots. Une sortie classique avec ce modèle est la matrice documents-termes.

Matrice documents-termes

La matrice documents-termes (**DTM**) utilise tous les mots de l'ensemble des données comme vocabulaire. C'est un objet matriciel mathématique faisant apparaître la fréquence des termes d'une collection de documents. Dans une DTM, chaque ligne de la matrice correspond à un

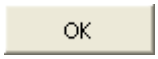
document et chaque colonne correspond à un terme (mot) dans le document. Chaque cellule représente la fréquence (nombre d'occurrences) du mot correspondant dans le document correspondant. Avant cela, une étape dite de « tokenisation » est effectuée afin d'extraire les mots de chaque document en utilisant le caractère espace comme séparateur. En plus de cela, de nombreuses options de filtrage peuvent être appliquées telles que la suppression des mots n'ayant pas d'importance significative dans la construction de la matrice. Ce prétraitement est appelé suppression des mots vides tels que : un, le, et, etc.). D'autres procédures de filtrage peuvent être effectuées comme la suppression des termes nuls (termes non présents au-dessus d'une certaine proportion sur l'ensemble des documents), ou bien la désuffixation des mots (les réduisant chacun à leur radical).

Matrice termes-documents

Dans une matrice fréquentielle termes-documents (**TDM**), chaque ligne de la matrice correspond à un terme (mot) et chaque colonne correspond à un document. Chaque cellule représente la fréquence (nombre d'occurrences) du mot correspondant dans le document correspondant.

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.


: cliquez sur ce bouton pour changer la façon dont vous chargez les données dans XLSTAT. Si le bouton possède une icône de souris, XLSTAT vous permet de faire une sélection à la souris des données. Si le bouton possède une icône de liste, XLSTAT vous permet de sélectionner les données à partir d'une liste. Si le bouton possède une icône orange, des boutons complémentaires avec un point d'interrogation  sont affichés pour vous permettre d'importer les données à partir de fichiers.

Onglet **Général** :

Feuille de calcul : sélectionnez une table composée de N documents (un document par cellule) sur la feuille Excel. Si l'option "Libellés des documents" a été sélectionnée, vérifiez que les cellules contenant des libellés ont été sélectionnées.

Fichiers (.txt) : sélectionnez plusieurs fichiers textes (version WINDOWS) ou le dossier qui les contient (version MAC). Chaque fichier est lu comme un document.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des colonnes : activez cette option si la première ligne des données sélectionnées (les données provenant de la feuille de calcul, les libellés des documents) contient un libellé.

Libellés des documents : activez cette option si vous voulez utiliser des libellés pour l'affichage des résultats. Si l'option « Libellés des colonnes » est activée, la première cellule de la sélection doit comprendre un en-tête. Si vous n'activez pas cette option, des libellés seront automatiquement créés (Doc 1, Doc 2, etc.).

Onglet **Options** :

Sous-onglet **Prétraitement** :

Filtrage de vocabulaire

Liste mots vides : activez cette option pour exclure une liste de mots vide (mots non significatifs figurant dans un texte) contenu dans les documents (langue anglaise par défaut).

Supprimer la ponctuation : activez cette option pour supprimer la ponctuation dans les documents.

Supprimer les nombres : activez cette option pour supprimer les nombres dans les documents.

Normalisation du texte

Racinisation : activez cette option pour réduire les mots à leur racine commune (racinisateur anglais par défaut).

Sous-onglet **Forme intermédiaire** :

Filtrage des termes

Supprimer les termes nuls : activez cette option pour supprimer les termes dont la proportion de présence est inférieure à 100 x (1- *valeur*)% sur l'ensemble des documents (*valeur* vaut 0.95 par défaut).

Fréquence minimum : activez cette option pour ignorer les termes apparaissant moins de *valeur* fois sur l'ensemble des documents (*valeur* vaut 2 par défaut).

Fréquence maximale : activez cette option pour ignorer les termes apparaissant plus de *valeur* fois sur l'ensemble des documents (*valeur* doit être supérieure ou égale au paramètre *Fréquence minimum* si celui-ci est activé).

Nombre maximum : activez cette option pour définir un nombre maximal de mots à inclure dans la matrice documents-termes (ou termes-documents). Les termes les moins fréquents seront supprimés.

Onglet **Sorties** :

Matrice termes-documents : activez cette option pour afficher la matrice termes-documents dans la feuille de résultats XLSTAT.

Matrice documents-termes : activez cette option pour afficher la matrice documents-termes (DTM) dans la feuille de résultats XLSTAT.

Exporter la matrice documents-termes (DTM) ou la matrice termes-documents (TDM) : activez cette option pour spécifier un chemin de dossier dans lequel exporter la matrice termes-documents ou la matrice documents-termes au format CSV (*comma-separated values*).

Onglet **Graphiques** :

Nuage de mots : activez cette option pour afficher un nuage de mots représentant tous les documents.

Résultats

La DTM ou la TDM est affichée. La matrice DTM ou TDM exportée (si l'option correspondante est choisie) n'a aucune limitation concernant le nombre maximum de termes autorisés à contenir (utile en cas de dépassement de la limitation du nombre de colonnes imposé par Excel dans la feuille de résultats).

À la fin de la matrice termes-documents, vous trouverez le bouton suivant :  . Ce bouton vous permet d'ouvrir automatiquement la boite de dialogue pré-remplie du Nuage de mot afin de créer et personnaliser de(s) nuage(s) de mots.

Exemple

Un exemple portant sur des données provenant du site *Internet Movie Database (IMBD)* est disponible en permanence sur le Centre d'aide XLSTAT :

https://help.xlstat.com/customer/fr/portal/articles/2937383-feature-extraction-tutorial-in-excel?b_id=9283

Bibliographie

Lewis, D. 1992 Text Representation for Text Classification.

Martin F Porter. 1980 An algorithm for suffix stripping.

David A Hull et al. 1996 Stemming algorithms: A case study for detailed evaluation. *JASIS 47*, 1 (1996), 70–84.

Analyse Sémantique Latente (LSA)

Utilisez l'Analyse Sémantique Latente ou Latent Semantic Analysis (LSA) afin de découvrir la sémantique cachée et sous-jacente (latente) de mots dans un corpus de documents.

Dans cette section :

Description

L'Analyse Sémantique Latente (LSA) permet de découvrir la sémantique cachée et sous-jacente (latente) de mots dans un corpus de documents en construisant des "concepts" liés aux documents et aux termes. La LSA utilise une matrice documents-termes en entrée qui décrit l'occurrence de certains termes dans les documents. C'est une matrice creuse dont les lignes correspondent aux "documents" et dont les colonnes correspondent aux "termes".

Utilisations de la LSA

Il existe plusieurs applications pour la LSA, parmi lesquelles :

- la comparaison de documents dans l'espace des concepts (classification et catégorisation de documents, partitionnement de données) ;
- la recherche de documents similaires entre différentes langues, en ayant accès à un dictionnaire de documents multilingues ;
- la recherche de relations entre les termes (résolution de synonymie et de polysémie) ;

Principe de la LSA

L'Analyse Sémantique Latente (LSA) s'appuie sur une matrice documents-termes. Les éléments de cette matrice contiennent les occurrences des différents termes dans chaque document.

Cette matrice est ensuite utilisée pour réaliser des associations entre les documents et des concepts (à partir des termes), et donc de relier les documents entre eux sur le plan sémantique

(une sorte de proximité thématique). Pour ce faire, on réalise différentes opérations mathématiques sur la matrice, dans l'ordre suivant :

- Calcul de la décomposition en valeurs singulières (SVD) de la matrice documents-termes M pour faire apparaitre les espaces propres des documents et des termes pour obtenir $M = D \times S \times T^T$
- Sélection des k premières valeurs singulières (par ordre d'importance) de la matrice diagonale S ainsi que les colonnes correspondantes dans les matrices T et D. La matrice D représente les vecteurs documents dans le nouvel espace des termes et la matrice T les vecteurs termes dans le nouvel espace des documents. Ci dessous l'expression de la version réduite de la SVD :

$$M_k = D_k \times S_k \times T_k^T$$

On peut ensuite trouver des termes similaires en calculant la similarité cosinus entre deux colonnes de la matrice de rang réduit M_k ce qui est strictement équivalant à la similarité cosinus entre les colonnes correspondantes de $S_k \times T_k^T$.

Le même principe s'applique pour trouver des documents similaires en calculant la similarité cosinus entre deux lignes de la matrice de rang réduit M_k ce qui est strictement équivalent à la similarité cosinus entre les lignes correspondantes de $D_k \times S_k$.

$$sim_{cosinus}(a_i, a_j) = \frac{a_i a_j^T}{||a_i||||a_j||}$$

avec $\{a_i, a_j\}$ pouvant être une paire de termes ou de documents.

Interprétation des résultats

La représentation des termes dans l'espace des k thématiques permet d'interpréter visuellement les similarités entre les termes d'une part, et entre les documents d'autre part.

En effet, qu'il s'agisse de la représentation des documents ou des termes dans l'espace des thématiques latentes, deux points très éloignés dans l'espace de la matrice d'origine peuvent apparaître proches dans un espace réduit à k dimensions latentes car la diminuation du rang a pour effet de fusionner les dimensions associées à des termes/documents ayant une signification similaire.

Nombre de thématiques

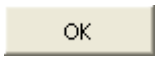
Deux méthodes sont communément utilisées pour déteminer quel nombre de facteurs doit être retenu pour l'interprétation des résultats :

Le *scree test* (Cattell, 1966) est fondé sur la courbe décroissante des valeurs propres. Le nombre de thématiques à retenir correspond au premier point d'inflexion détecté sur la courbe.

On peut aussi se fonder sur le pourcentage cumulé de variabilité représenté par les axes factoriels et décider de se contenter d'un certain pourcentage.

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.


: cliquez sur ce bouton pour changer la façon dont vous chargez les données dans XLSTAT. Si le bouton possède une icône de souris, XLSTAT vous permet de faire une sélection à la souris des données. Si le bouton possède une icône de liste, XLSTAT vous permet de sélectionner les données à partir d'une liste. Si le bouton possède une icône orange, des boutons complémentaires avec un point d'interrogation  sont affichés pour vous permettre d'importer les données à partir de fichiers.

Onglet **Général**:

Matrice documents/termes : sélectionnez un tableau comprenant N documents décrits par P termes. Si des en-têtes de colonnes ont été sélectionnés, veuillez vérifier que l'option "Libellés des termes" est activée.

Poids des documents : activez cette option si vous voulez pondérer les documents. Si vous n'activez pas cette option, les poids seront tous considérés comme valant 1. Ces derniers doivent être impérativement supérieurs ou égaux à 0. Si l'option "Libellés des termes" est activée, la première cellule de la sélection doit comprendre un en-tête.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des termes : activez cette option si la première ligne des données sélectionnées (Matrice documents-termes, libellés des documents, poids des documents) contient un libellé.

Libellés des documents : activez cette option si vous voulez utiliser des libellés des documents pour l'affichage des résultats. Si l'option « Libellés des termes » est activée, la première cellule de la sélection doit comprendre un en-tête. Si vous n'activez pas cette option, des libellés seront automatiquement créés (Doc1, Doc2, …).

Onglet **Options**:

Nombre de thématiques : entrez le nombre de thématiques envisagé pour lequel l'Analyse Sémantique Latente sera appliquée.

Regrouper par document : activez cette option si vous voulez créer des classes de documents dans l'espace sémantique crée. Ces classes peuvent être affichées via l'option **Colorer par classe** en dessous de la case à cocher **Matrice de corrélation document-document** dans l'onglet graphique.

Regrouper par terme : activez cette option si vous voulez créer des classes de termes dans l'espace sémantique crée. Ces classes peuvent être affichées via l'option **Colorer par classe** en dessous de la case à cocher **Matrice de corrélation terme-terme** dans l'onglet graphique.

Type de classification : vous pouvez activer l'une des deux options suivantes afin de choisir le type de classification relativement aux deux options de regroupement expliqués ci-dessous :

- **Absolue** : choisissez cette option pour effectuer une classification dans le nouvel espace sémantique créé dans lequel chaque élément (terme / document) ne peut appartenir qu'à une seule thématique à la fois pour représenter une classe (hard clustering).

- **Floue** : choisissez cette option pour effectuer une classification dans le nouvel espace sémantique créé dans lequel chaque élément (terme / document) peut appartenir à plusieurs thématiques à la fois pour représenter une classe(Soft clustering en anglais).

Conditions d'arrêt :

- **Itérations** : entrez le nombre maximal d'itérations pour l'algorithme SVD. Les calculs sont interrompus dès que le nombre maximal d'itérations est dépassé. Valeur par défaut : 500. Si la condition d'arrêt n'est pas sélectionnée, l'algorithme itérera jusqu'au rang maximal de la matrice documents-termes d'entrée.

Onglet **Prétraitement** :

Données manquantes :

Ne pas accepter les valeurs manquantes : activez cette option pour que XLSTAT empêche la poursuite des calculs si des valeurs manquantes sont détectées.

Supprimer les observations : activez cette option pour supprimer les observations comportant des données manquantes.

Remplacer les valeurs par 0 : activez cette option pour remplacer les valeurs manquantes par 0.

Onglet **Sorties**:

Tableau de synthèse : activez cette option pour afficher la synthèse de l'Analyse Sémantique Latente. Ceci inclut une énumération du nombre de termes et documents pour chacune des thématiques ainsi que le tableau (*scree plot*) des valeurs propres liées aux thématiques latentes issues de la décomposition. Les valeurs sont affichées dans un ordre décroissant d'amplitude et de variabilité expliquée.

Tableau des thématiques : activez cette option pour afficher le tableau des termes composant chaque thématique.

- **Max. termes/thématique** : activez cette option afin de spécifier le nombre de termes au maximum à afficher dans la tableau des thématiques. Cette valeur sera également appliquée sur l'affichage des matrices de corrélations dans les graphiques.

Termes les plus similaires : activez cette option pour afficher le tableau des termes les plus similaires pour un terme donné dans l'espace sémantique crée.

- **Nombre de termes** : activez cette option afin de spécifier le nombre de termes sur lesquels calculer les termes les plus similaires.
- **Nombre de termes voisins** : activez cette option afin de spécifier le nombre de plus proches voisins à afficher dans le tableau des termes les plus similaires. Ces derniers seront affichés de gauche à droite par ordre décroissant de similarité.

Onglet **Graphiques** :

Scree plot : activer cette option pour afficher le graphique (*scree plot*) des valeurs propres liées aux thématiques latentes issues de la décomposition. Les valeurs sont affichées dans un ordre décroissant d'amplitude et de variabilité expliquée.

Matrice de corrélation terme-terme : activez cette option pour afficher la matrice de corrélation représentant les corrélations (similarités) terme-terme dans le nouvel espace sémantique.

- **Colorer par classe** : activez cette option pour colorer les termes en relation avec chaque classe (thématique) issue de la classification en amont.

Matrice de corrélation document-document : activez cette option pour afficher la matrice de corrélation représentant les corrélations (similarités) document-document dans le nouvel espace sémantique.

- **Colorer par classe** : activez cette option pour colorer les documents en relation avec chaque classe (thématique) issue de la classification en amont.

Legende : activez cette option pour que la légende des différentes matrices de corrélation soit affichée sur les graphiques. Cette dernière sera cependant non disponible lorsque l'option *Colorer par classe* est activée.

Résultats

Tableau de synthèse : ce tableau de synthèse présente pour chacune des thématiques, le nombre total de termes-documents les composant. L'utilisateur a ainsi la possibilité par la suite d'afficher l'intégralité de ces derniers dans les graphiques liés aux *matrices de corrélation* ainsi que dans le *tableau des thématiques*.

Les valeurs propres et le graphique (*scree plot*) correspondant sont aussi affichés. La variance cumulée fournit une indication sur la pertinance des thématiques calculées. Plus cette dernière est élévée, meilleure est l'approximation issue de la SVD "tronquée".

Tableau des thématiques : ce tableau affiche la liste des termes par thématique de gauche à droite par ordre décroissant de relation avec la thématique concernée.

Termes les plus similaires : ce tableau affiche les n termes les plus similaires au terme sélectionné dans la liste déroulante dans le nouvel espace sémantique crée et ce par ordre décroissant de similarité.

Matrices de corrélation : ces graphiques des corrélations (terme-terme, document-document, terme-document) permettent de visualiser le degré de similarité (similarité cosinus) entre les termes (**Matrice de corrélation terme-terme**) ou bien les documents (**Matrice de corrélation document-document**) ou entre les termes et documents (**Matrice de corrélation terme-document**) dans leurs espaces respectifs. Les similarités sont comprises entre 0 et 1, la valeur 1 correspondant à une similarité parfaite dans les deux sens (positif et négatif).

Exemple

Un exemple d'utilisation de l'Analyse Sémantique Latente est disponible sur le Centre d'aide XLSTAT à l'adresse

https://www.xlstat.com/demo-lsaf.htm

Bibliographie

Landauer, T.K., & Dumais, S.T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge.

Berry M.W. (1994). Computing the Sparse Singular Value Decomposition via SVDPACK. In *Recent Advances in Iterative Methods*, IMA Volumes in Mathematics and its Application, **60**, Springer, New York, 13-29.

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research*, **1**, 245-276.

Analyse de sentiment

Utilisez cet outil pour déterminer l'opinion d'un document en anglais grâce à la librairie Syuzhet.

Dans cette section :

Description

L'analyse de sentiment est le processus d'extraction de l'intention émotionnelle d'un auteur à partir d'un texte (Ted Kwarler, 2017). Elle permet de catégoriser un commentaire, un livre ou un document en général. Un document peut être catégorisé comme une opinion positive, négative ou neutre.

Quand utiliser l'analyse de sentiment ?

L'analyse de sentiment aide les entreprises à comprendre les critiques et les retours des clients, les critiques des produits, à analyser les commentaires sur les sites web (comme les tweets ou posts) ou encore les discussions politiques. En général, l'analyse de sentiment répond à la question : "Comment les personnes (clients) se sentent-ils par rapport à quelque chose ?".

Qu'utilise l'analyse de sentiment ?

L'analyse de sentiment utilise un dictionnaire où les termes ont un score ou sont catégorisés (positif, négatif ou neutre). Chaque dictionnaire utilise sa propre échelle c'est pourquoi XLSTAT propose quatre dictionnaires de sentiment afin d'assigner un sentiment à chaque terme :

- Analyse de sentiment avec le dictionnaire de **Bing** : 6789 termes anglais sont labellisés comme "négatif", "neutre" ou "positif" dans le dictionnaire de Bing. Un terme catégorisé comme "négatif" obtient un score de -1, s'il est catégorisé comme "neutre" il obtient un score de 0 enfin il obtient 1 lorsqu'il est catégorisé en "positif".

- Analyse de sentiment avec le dictionnaire de **Syuzhet** : 10748 termes anglais sont notés entre -1 et 1 dans le dictionnaire de Syuzhet. Un terme est dit "négatif" lorsque son score est inférieur à 0, et au contraire il est dit "positif" lorsque son score est supérieur à 0.

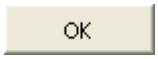
- Analyse de sentiment avec le dictionnaire de **AFINN** : 3382 termes anglais sont notés entre -5 et 5 dans le dictionnaire d'AFINN. Un terme est dit "négatif" lorsque son score est inférieur à 0, et au contraire il est dit "positif" lorsque son score est supérieur à 0.

- Analyse de sentiment avec le dictionnaire **NRC** (échelle d'émotions) : ce dictionnaire catégorise 13901 termes anglais avec huit émotions basiques (colère, peur, anticipation, confiance, surprise, tristesse, joie et dégoût) et deux sentiments (négatif ou positif).

En plus d'un dictionnaire de sentiment, l'analyse de sentiment a besoin de documents qui ont subi une tokenisation. XLSTAT propose d'utiliser en amont l'outil d'Extraction de caractéristique afin d'obtenir la matrice documents-termes.

Comment est calculé le score du document ? Le score de chaque terme présent dans un document est multiplié par sa fréquence puis les scores sont sommés afin d'obtenir le score du document.

Boîte de dialogue



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.


: cliquez sur ce bouton pour changer la façon dont vous chargez les données dans XLSTAT. Si le bouton possède une icône de souris, XLSTAT vous permet de faire une sélection à la souris des données. Si le bouton possède une icône de liste, XLSTAT vous permet de sélectionner les données à partir d'une liste. Si le bouton possède une icône orange, des boutons complémentaires avec un point d'interrogation  sont affichés pour vous permettre d'importer les données à partir de fichiers.

Onglet **Général** :

Fréquences des termes : sélectionnez les fréquences des termes dont une colonne correspond aux fréquences d'un terme dans chaque document. Si l'option « Libellés des colonnes » est activée, la première cellule de la sélection doit comprendre un en-tête.

Dictionnaire de sentiment : choisissez parmi quatre dictionnaires de sentiment (voir la description).

Scores personnalisés : sélectionnez deux colonnes contenant un terme et son score. Si vous utilisez le dictionnaire de Bing, vous devez entrer "negative", "neutral" ou "positive". Cette option permet de définir le sentiment d'un terme indépendamment du dictionnaire choisi précédemment. Si l'option « Libellés des colonnes » est activée, la première cellule de la sélection doit comprendre un en-tête. Pour cette sélection, les données manquantes sont lues comme "neutre" ou zéro. Remarque : non disponible pour le dictionnaire NRC.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des colonnes : activez cette option si la première ligne des données sélectionnées contient un libellé.

Libellés des documents : activez cette option si vous voulez utiliser des libellés de document pour l'affichage des résultats. Si l'option « Libellés des colonnes » est activée, la première cellule de la sélection doit comprendre un en-tête. Si vous n'activez pas cette option, des libellés seront automatiquement créés (Doc1, Doc2, etc.).

Onglet **Données manquantes** :

Supprimer les observations : activez cette option pour supprimer les observations comportant des données manquantes.

Ignorer les données manquantes : activez cette option pour ignorer les données manquantes.

Onglet **Sorties** :

Fréquences et scores des termes : activez cette option pour afficher un tableau montrant la fréquence totale et le score de chaque terme inclus dans la sélection de fréquence des termes. Remarque : non disponible pour le dictionnaire NRC.

Fréquences et émotion(s) associée(s) des termes : activez cette option pour afficher un tableau montrant la fréquence totale et les émotions associées à chaque terme inclus dans la sélection de fréquence des termes. Remarque : seulement disponible pour le dictionnaire NRC.

- **Afficher les termes ayant un sentiment seulement** : activez cette option pour afficher les termes ayant un sentiment seulement. Les termes avec un sentiment neutre, ce qui signifie que leur score est égal à zéro ou qu'ils ne sont associés à aucune émotion, ne sont pas affichés.

Fréquences globales des émotions : activez cette option pour afficher la fréquence totale de chaque émotion présente dans tous les documents. Remarque : seulement disponible pour le dictionnaire NRC.

Scores des documents : activez cette option pour afficher un tableau montrant le score de chaque document selon le dictionnaire de sentiment choisi dans l'onglet Général. Remarque : non disponible pour le dictionnaire NRC.

- **Trier par score (décroissant)** : activez cette option pour trier les scores des documents dans l'ordre décroissant.

Fréquences des émotions par document : activez cette option pour afficher un tableau indiquant la fréquence de chaque émotion dans chaque document. Remarque : seulement disponible pour le dictionnaire NRC.

Interprétation du résultat : activez cette option pour afficher sous les tableaux de résultats, une courte interprétation.

Onglet **Graphiques** :

Fréquence des termes : activez cette option pour afficher un diagramme en barre montrant la fréquence totale des termes.

- **Fréquence minimum** : entrez la fréquence minimum qu'un terme doit avoir pour être affiché dans le graphique. Nous vous suggérons d'augmenter la fréquence minimum lorsque le nombre de termes augmente.

Scores des termes : activez cette option pour afficher un diagramme en barre montrant les scores des termes.

Score des documents : activez cette option pour afficher un diagramme en barre montrant les scores des documents. Si l'option **Trier par score (décroissant)** est activée alors le diagramme est aussi trié.

Distribution des scores des documents : activez cette option pour afficher un histogramme montrant la distribution des scores des documents.

Fréquences globales des émotions : activez cette option pour afficher un diagramme en barre montrant la fréquence totale des émotions. Remarque : seulement disponible pour le dictionnaire NRC.

Nuage de mots basé sur les sentiments : activez cette option pour afficher un nuage de mot où les termes sont colorés selon leur sentiment (positif, négatif ou les émotions associées).

- **Termes maximums** : entrez le nombre maximum de termes à inclure dans le nuage de mot basé sur les sentiments.

Interprétation du résultat : activez cette option pour afficher sous les graphiques une courte interprétation.

Résultats

Résultats associés aux scores des documents : le tableau et le graphique associés aux scores des documents sont affichés pour donner un aperçu du sentiment de chaque document selon l'échelle du dictionnaire de sentiment utilisé. Si l'option **Trier par score (décroissant)** n'est pas activée, vous pouvez voir l'évolution du score des documents, spécialement s'ils sont entrés dans un ordre chronologique.

Résultats associés aux documents et les émotions associées : avec l'échelle des émotions (NRC), un tableau est affiché pour montrer les fréquences de chaque émotion dans un document. Ce tableau peut être complété avec les scores des documents obtenus avec un autre dictionnaire de sentiment, ce qui permet de mettre des mots naturels sur le sentiment ou l'intensité d'une opinion présente dans un document.

Résultat associé à la distribution des scores des documents : l'histogramme affiché aide à connaître la fréquence des scores. Dans le cas où les scores sont centrés en 0, cela signifie que les documents ont en moyenne une majorité de mots neutres. D'un autre côté, si les scores

sont centrés en une valeur supérieure (resp. inférieure) à 0, cela signifie que les documents ont en moyenne au moins un mot positif (resp. négatif).

Résultats associés aux fréquences des termes : le tableau et graphique associés aux fréquences des termes sont affichés pour donner un aperçu de la fréquence totale des termes, en d'autres termes cela montre le nombre d'occurrences d'un terme parmi tous les documents.

Résultats associés aux scores des termes : le tableau et graphique associés aux scores des termes sont affichés pour donner un aperçu du sentiment de chaque terme selon l'échelle du dictionnaire de sentiment. Avec l'échelle d'émotions, un terme peut être associé à zéro, un ou plusieurs termes. Les termes neutres ont une cellule vide dans la colonne "Score". Les scores personnalisés sont affichés en gras.

Exemple

Un tutoriel sur la façon d'utiliser l'analyse de sentiment est disponible sur le Centre d'aide XLSTAT :

https://www.xlstat.com/demo/stm_en

Bibliographie

Kwartler, T. (2017). Text mining in practice with R. John Wiley & Sons.

Jockers, M. (2017). Package ‘syuzhet’. URL: https://cran. r-project. org/web/packages/syuzhet.

Mejova, Y. (2009). Sentiment analysis: An overview. *University of Iowa, Computer Science Department*.

Sélection de termes

Utilisez cette méthode pour faire une régression sur une matrice documents-termes.

Dans cette section :

Description

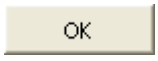
La sélection de termes utilise la très célèbre méthode de régression Elastic-net ainsi que sa version logistique. En effet, cela permet de modéliser des variables quantitatives mais aussi binomiales (typiquement binaires) et multinomiales (variables qualitatives à plus de deux modalités).

La sélection de termes est une méthode utilisée dans le cas du text mining, où la matrice documents-termes remplace les variables quantitatives explicatives, et le vecteur de sentiment est la variable réponse donnant le sentiment ("positif", "négatif", etc.) de chaque document ou sa note (indication quantitative de l'opinion).

La régression Elastic-net est basée sur deux paramètres fondamentaux : le paramètre de compromis α (compris entre 0 et 1) et le paramètre de régularisation $\lambda > 0$. XLSTAT offre la possibilité de trouver le λ optimal par validation croisée.

Pour en savoir plus sur la régression Elastic-net, sa description est disponible ici. Pour la régression Elastic-net logistique, la théorie de la méthode Elastic-net est à compléter avec la théorie de la régression logistique.

Boîte de dialogue



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.

 : cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.

 : cliquez sur ce bouton pour changer la façon dont vous chargez les données dans XLSTAT. Si le bouton possède une icône de souris, XLSTAT vous permet de faire une sélection à la souris des données. Si le bouton possède une icône de liste, XLSTAT vous permet de sélectionner les données à partir d'une liste. Si le bouton possède une icône orange, des boutons complémentaires avec un point d'interrogation  sont affichés pour vous permettre d'importer les données à partir de fichiers.

Onglet **Général** :

Variable réponse : sélectionnez la variable réponse à modéliser. Si l'option « Libellés des colonnes » est activée, la première cellule de la sélection doit comprendre un en-tête.

Type de réponse : sélectionner le type de la variable réponse.

- **Gaussien** : si votre variable réponse est quantitative, choisissez ce type pour ajuster un modèle de régression ;
- **Poisson** : si votre variable réponse est quantitative, choisissez ce type pour ajuster un modèle de régression ;
- **Binomiale** : si votre variable réponse est binaire, choisissez ce type pour ajuster un modèle de régression logistique ;
- **Multinomiale** : si votre variable réponse comporte plus de deux catégories, choisissez ce type pour ajuster un modèle de régression logistique.

Fréquences des termes : sélectionnez les fréquences des termes dont une colonne correspond aux fréquences d'un terme dans chaque document. Les données doivent être quantitative. Si l'option « Libellés des colonnes » est activée, la première cellule de la sélection doit comprendre un en-tête.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des colonnes : activez cette option si la première ligne des données sélectionnées contient un libellé.

Libellés des documents : activez cette option si vous voulez utiliser des libellés de document pour l'affichage des résultats. Si l'option « Libellés des colonnes » est activée, la première cellule de la sélection doit comprendre un en-tête. Si vous n'activez pas cette option, des libellés seront automatiquement créés (Doc1, Doc2, etc.).

Onglet **Options** :

Alpha : α correspond au paramètre compromis compris entre 0 et 1. Quand $\alpha = 1$, c'est la pénalité LASSO qui est appliquée, et quand $\alpha = 0$ c'est la pénalité Ridge.

Lambda : choisissez les valeurs de λ à tester pendant la validation croisée.

- **Automatique** : sélectionnez cette option pour générer automatiquement des valeurs de λ .
  - **Nombre de valeurs de lambda** : entrez le nombre de valeurs de λ à générer. Valeur par défaut : 100.
- **Valeurs de lambda personnalisées** : sélectionnez cette option pour entrer manuellement les valeurs de λ en sélectionnant une colonne contenant autant de lignes que de valeurs de λ à tester.
- **Itérations** : entrez le nombre maximal d'itérations. Les calculs sont interrompus dès que le nombre maximal d'itérations est dépassé. Valeur par défaut : 10000.

Nombre de blocs : entrez le nombre de blocs à constituer pour la validation croisée. Valeur par défaut : 10.

Maximum variables : entrez le nombre maximum de variables à utiliser dans le modèle.

Onglet **Prédiction** :

Fréquence de termes (Prédiction) : activez cette option si vous souhaitez sélectionner des données à utiliser en mode prédiction. Si vous activez cette option, vous devez veiller à ce que les données de prédiction soient organisées comme les données d'estimation : mêmes variables, même ordre dans les sélections. Si l'option « Libellés des colonnes » est activée, la première cellule de la sélection doit comprendre un en-tête.

Libellés des documents (Prédiction) : activez cette option si vous voulez utiliser des libellés d'observations disponibles sur une feuille Excel pour l'affichage des résultats. Si l'option « Libellés des colonnes » est activée, la première cellule de la sélection doit comprendre un en-tête. Si vous n'activez pas cette option, des libellés seront automatiquement créés (PredDoc1, PredDoc2, …).

Onglet **Données manquantes** :

Ne pas accepter les données manquantes : activez cette option si vous voulez que les calculs soient interrompus et que vous soyez prévenu en cas de présence de données manquantes.

Supprimer les observations : activez cette option pour supprimer les observations comportant des données manquantes.

Onglet **Sorties** :

Sélectionnez les coefficients selon le : sélectionnez les coefficients selon le λ optimal de votre choix. * **Lambda minimum** : sélectionnez cette option pour choisir les coefficients associés au λ qui donne la plus petite erreur moyenne de validation croisée ; * **Lambda 1se** :

sélectionnez cette option pour choisir les coefficients associés au λ qui donne la plus petite erreur moyenne de validation croisée ; * **Lambda 1se** : sélectionnez cette option pour choisir les coefficients associés au λ donnant le modèle le plus régularisé tel que l'erreur de validation croisée est à un écart-type du minimum.

Lambda optimaux : activez cette option pour afficher un tableau donnant les valeurs et les degrés de liberté associés aux λ.

Coefficients : activez cette option pour afficher les coefficients triés de chaque terme.

Odds ratio : activez cette option pour afficher les odds ratio de chaque terme dans le même tableau que les coefficients.

Fréquences des termes : activez cette option pour afficher la fréquence totale de chaque terme dans le même tableau que les coefficients.

Afficher les coefficients non-nuls seulement : activez cette option pour afficher seulement les termes avec une influence sur le modèle. Les termes avec des coefficients nuls, leur odd ratio et leur fréquence sont supprimés du tableau "Résultats par terme".

Résultats par document : activez cette option pour afficher la variable réponse et la prédiction pour chaque document et les probabilités pour la classification.

Matrice de confusion : cette option est uniquement disponible pour la classification. Elle permet d'afficher la matrice de confusion des résultats de prédiction sur l'échantillon d'apprentissage. La matrice de confusion contient les informations concernant les classifications observées et prédites par l'algorithme. Les performances de l'algorithme peuvent être évaluées au moyen de cette matrice de confusion. La diagonale contient les prédictions correctes. Plus la somme des éléments de la diagonale est importante, meilleur est le classifieur.

Coefficients d'ajustement : dans ce tableau sont affichées les statistiques relatives à l'ajustement du modèle de régression.

Onglet **Graphiques** :

Coefficients : activez cette option pour afficher un diagramme en barre montrant les coefficients associés à chaque terme.

Odds ratio : activez cette option pour afficher un diagramme en barre montrant les odds ratio.

Évolution de la déviance : activez cette option pour afficher un graphique montrant la courbe de validation croisée avec ses courbes d'écart-type supérieure et inférieure, en tant que fonction des valeurs de λ automatiquement générées ou entrées manuellement (voir l'onglet Options). Le λ minimum est tracé en rouge tandis que λ 1se est tracé en bleu sur ce graphique. Si les deux λ sont égaux, seul le λ minimum est tracé.

Résultats

Résultats associés aux termes : ce tableau donne un aperçu de l'influence de chaque terme. Les coefficient et odd ratio permettent de savoir si un terme est important ou non dans le modèle. Le coefficient donne l'intensité et la direction de l'influence tandis que l'odd ratio donne la probabilité de prédire la classe cible vs une autre valeur. Par exemple, si la classe cible est

"Positive" et la deuxième est "Négative" et que l'odd ratio pour le terme "good" est 3, cela signifie que le document qui contient "good" aura trois fois plus de chance d'être prédit "Positive" qu'un document n'ayant pas ce terme. La colonne des fréquences aide à savoir si le coefficient est influencé par une fréquence élevée. Si tous les coefficients sont nuls, la constante est la seule à être affichée sur les graphiques. Pour avoir plus de coefficients non-nuls, nous vous suggérons de diminuer la valeur d'α.

Résultats associés aux matrices de confusion : Les matrices de confusion sont déduites des classifications obtenues et de la classe effective ainsi que les pourcentages d'observations correctement classifiées.

Résultats associés aux coefficients d'ajustement : dans ce tableau sont affichées les statistiques relatives à l'ajustement du modèle de régression :

- **Observations** : le nombre d'observations prises en compte dans les calculs ;
- **DDL** : le nombre de degrés de liberté pour le modèle retenu (ce nombre est égal au nombre de coefficients non-nuls dans le modèle) ;
- **Déviance** : correspond à la perte, pour le modèle Gaussien il s'agit de l'erreur au carré, pour le modèle de Poisson il s'agit de la déviance et pour la classification binomiale ou multinomiale il s'agit de l'erreur de classification ;
- **AIC** : le critère d'information d'Akaike (Akaike's Information Criterion) ;
- **AICc** : le critère d'information d'Akaike corrigé (Corrected Akaike's Information Criterion) ;
- **SBC** : le critère bayésien de Schwarz (Schwarz's Bayesian Criterion).

Résultats associés aux documents : ce tableau donne un aperçu des prédictions des documents. Pour le cas de la classification, les probabilités de la classe cible sont affichées pour la classification binomiale et les probabilités de toutes les classes sont affichées pour la classification multinomiale. Remarque : la classe cible est la dernière classe dans l'ordre alphabétique.

Exemple

Un tutoriel sur la façon d'utiliser la sélection de termes est disponible sur le Centre d'aide XLSTAT :

https://www.xlstat.com/demo/trs_en

Bibliographie

Hastie, T., Qian, J., & Tay, K. (2021). An Introduction to glmnet. CRAN R Repository.

Aide à la décision

Aide Multicritère à la décision : méthodes ELECTRE

Dans un processus décisionnel, les méthodes ELECTRE aident à identifier un ensemble de solutions à un problème donné, à les comparer entre elles ou bien à les classer de la meilleure à la moins bonne. Ces méthodes ont l'avantage de prendre en compte plusieurs critères qui peuvent être de nature différente (qualitatif ou quantitatif) et dont l'ordre d'importance peut être choisi.

Dans cette section :

Description

Les méthodes ELECTRE, dont l'acronyme désigne ELicitation Et Choix Traduisant la REalité, regroupent une famille de méthodes d'aide à la décision dont la particularité est l'agrégation partielle via la construction de relations de comparaisons des performances de chaque couple de solutions. Contrairement aux méthodes d'optimisation classique qui consistent à formuler le problème sous la forme d'une fonction coût et à rechercher son optimum, ici on compare les solutions 2 à 2, critère par critère mettant ainsi en avant une préférence/indifférence d'une réponse par rapport à une autre et aboutissant à une matrice de surclassement. Ces méthodes ont l'avantage d'accepter des situations d'incomparabilité aux critères parfois qualitatifs et incommensurables.

On appelle actions potentielles, ou encore alternatives, les différentes solutions offertes à un problème décisionnel. Ces actions sont listées de façon exhaustive ou non et doivent être formulées par l'utilisateur. Les conséquences de chacune d'elle sont évaluées à l'aide de critères. Un critère peut être qualitatif ou quantitatif et doit être défini par l'utilisateur. Lorsqu'il est qualitatif, l'évaluation des actions sur ce critère doit être ramenée à une échelle numérique définie par l'utilisateur. Par exemple soit le critère « type de diplôme » pour la sélection d'un candidat à un recrutement. Ce critère va être ramené à une échelle numérique arbitraire qui peut être : 0 pour baccalauréat, 1 pour BTS, 2 pour ingénieur, etc … Pour permettre une contribution différente de ces critères dans le problème décisionnel, l'utilisateur peut fournir un poids à chacun qui augmente avec l'importance de celui-ci. Dans l'exemple de la sélection d'un candidat ajoutons un nouveau critère « l'âge du candidat » allant de 25 à 50 ans. En supposant que ce critère soit moins important que le critère « type de diplôme », alors la valeur 1 sera

affectée au poids du critère « Age du candidat » et la valeur 2 au poids du critère « type de diplôme ». Au final, les informations qui doivent être fournies à minima par l'utilisateur pour se servir d'une méthode ELECTRE sont la liste des actions, la liste des critères, l'évaluation de chaque action par critère et le poids de chaque critère.

Soit A un ensemble fini de p actions potentielles, A = {a_1, a_2, …, a_p}. Soit F une famille cohérente de n critères, F = {g_1, g_2, …, g_n} dont chaque fonction g, représentant l'évaluation des actions par critère, est définie sur A et prend ses valeurs dans un ensemble totalement ordonné. Soit K l'ensemble de valeurs des poids associés à chaque critère, K = {k_1, k_2, …, k_p}. Le tableau dit des performances est alors composé de n lignes et p colonnes.

On dissocie un vrai-critère d'un pseudo-critère en fonction de l'exactitude des évaluations. Si celles-ci sont faciles à établir en étant dépourvues de marges d'erreurs, on parle de vrai-critère. A l'inverse si les évaluations sont floues et imprécises, on parle de pseudo-critère. Dans ce deuxième cas, la méthode de calcul sera complétée de seuils de discrimination afin d'accroitre le réalisme de la modélisation des préférences.

XLSTAT propose deux méthodes ELECTRE, 1 et 3. Une description de chacune d'elle est faite ci-dessous

Electre 1

Cette méthode est privilégiée pour identifier un ensemble de solutions à un problème décisionnel. Les critères sont des vrai-critères. Soient a et b deux actions potentielles, Electre 1 permet d'obtenir une matrice de surclassement traduisant numériquement les assertions « a surclasse b », noté aSb, c'est-à-dire que l'action a est privilégiée à l'action b ou l'assertion contraire. Pour ce faire, nous avons besoin de calculer deux matrices, la première appelée matrice de concordance et la deuxième matrice de discordance.

Les indices de la matrice de concordance pour deux actions a et b sont notés par C (a, b), compris entre 1 et 0, et mesurent la pertinence de l'assertion « a surclasse b » comme suit : $\forall a, b \in A$

$$C(a,b) = \frac{1}{\sum_{j=1}^{n} k_j} \sum_{j=1}^{n} k_{j/g_j(a) \geq g_j(b)} \qquad (1)$$

Les indices de la matrice de discordance sont notés D (a, b), compris entre 1 et 0, et mesurent la pertinence d'un argument en défaveur de l'assertion « a surclasse b », comme suit : $\forall a, b \in A$

$$D(a,b) = \frac{1}{\delta} \max_{j=1 \text{à} n} (g_j(b) - g_j(a)), \qquad (2)$$

Où

$$\delta = \max_{j=1 \text{à} n} (\max_{i=1 \text{à} p} (g_{j,i}(a) - g_{j,i}(b))).$$

La matrice de surclassement est construite à partir de l'ensemble de ces 2 indices via la relation de surclassement suivante : $\forall a, b \in A$

$$aSb \Leftrightarrow \begin{cases} C(a,b) \geq \hat{c} \\ D(a,b) \leq \hat{d} \end{cases} \quad (3)$$

où $\hat{c}$ désigne le seuil de concordance et $\hat{d}$ le seuil de discordance. Ces seuils doivent être pris dans l'intervalle [0,1]. Lorsque les 2 inégalités de (1) sont vraies alors l'indice de surclassement vaut 1, sinon il est égal à 0. On peut ainsi en déduire, pour chaque action, le nombre de fois qu'elle surclasse et le nombre de fois qu'elle est surclassée. Ce résultat est synthétisé dans un tableau classant les actions en fonction du nombre de surclassement. Les actions obtenant le même nombre sont classées au même rang. Par défaut les seuils sont fixés à respectivement 1 et 0 mais pour plus de souplesse dans la méthode et affaiblir l'assertion aSb ils peuvent être variés par l'utilisateur.

La méthode est complétée d'une analyse de sensibilité sur les seuils de concordance et de discordance. Ceci permet d'identifier les valeurs minimales et maximales pour lesquels le résultat final du surclassement reste inchangé. Pour ce faire, Electre 1 est calculée 4 fois : 2 fois en modifiant la valeur du seuil de concordance et en gardant la valeur du seuil de discordance à la valeur fournie par l'utilisateur (ou à la valeur par défaut 0) et 2 fois en gardant le seuil de concordance à la valeur donnée par l'utilisateur (ou à la valeur par défaut 1) et en modifiant le seuil de discordance. Les valeurs sont modifiées à plus ou moins 10% de la valeur prédéfinie.

Electre 3

Cette méthode est privilégiée pour classer un ensemble de solutions de la meilleure à la moins bonne. Les critères sont des pseudo-critères et dans ce cas des seuils sont requis pour faire l'analyse. Comparée à Electre 1, Electre 3 nécessite plus de calculs pour arriver au résultat. En effet, Electre 3 va d'abord calculer un score résumant l'information de concordance et de discordance entre les actions du problème. Ce score va être ensuite utilisé pour construire deux préclassements, un premier classant les actions de la meilleure à la moins bonne et un second classant de la moins bonne à la meilleure. La matrice de surclassement est déduite par croisement de ces 2 préclassements et avec l'aide du classement final.

On note q_j le seuil d'indifférence, p_j le seuil de préférence et v_j le seuil véto tels que $q_j < p_j < v_j$ sur le critère j. Soit u_j = $g_j(a) - g_j(b)$ la différence entre les performances de deux actions a et b sur le critère j. On compare u_j aux différents seuils et on définit les relations de surclassement suivantes :

- a et b sont indifférentes (a I b) $\Leftrightarrow u_j \leq q_j(g_j(a))$,
- a est faiblement préférée à b (a R b) $\Leftrightarrow q_j(g_j(a)) \leq u_j \leq p_j(g_j(a))$,
- a est strictement préférée à b (a P b) $\Leftrightarrow p_j(g_j(a)) \leq u_j$,
- a est moins bonne que l'action b (a NP b) $\Leftrightarrow u_j \geq v_j(g_j(a))$.

L'objectif d'Electre 3 est alors d'établir une matrice de surclassement traduisant les assertions aSb avec S = I, R, P et NP, pour tous les couples d'actions a et b de A. Pour ce faire, nous avons besoin de calculer les indices de crédibilité selon l'équation suivante :

$$d(a,b) = \begin{cases} C(a,b) & \text{si} \quad \forall j \quad D_j(a,b) > C(a,b), \\ C(a,b) \prod_{j=1}^{n} \frac{1-D_j(a,b)}{1-C(a,b)} \end{cases} \qquad (4)$$

Où C(a, b) désigne l'indice de la matrice de concordance globale et $D_j(a,b)$ l'indice de la matrice de discordance partielle j. Ces matrices sont respectivement calculées de la manière suivante : la concordance globale entre deux actions a et b est une combinaison linéaire des concordances partielles $c_j(a,b)$ liées à chaque critère j, et normalisée par le poids des critères, via l'équation :

$$C(a,b) = \frac{\sum_{j=1}^{n} k_j \times c_j(a,b)}{\sum_{j=1}^{n} k_j} \qquad (5)$$

avec

$$c_j(a,b) = \frac{p_j(g_j(a)) - min\left(g_j(b) - g_j(a), p_j(g_j(a))\right)}{p_j(g_j(a)) - min\left(g_j(b) - g_j(a), q_j(g_j(a))\right)}$$

Ce calcul fait intervenir la distance u_j, les seuils de préférence et d'indifférence. Le calcul des matrices de discordance partielle fait lui intervenir la distance u_j, les seuils de préférence et véto et est donné par l'équation suivante :

$$D_j(a,b) = Min\left(1; Max\left(0; \frac{g_j(b) - g_j(a) - p_j(g_j(a))}{v_j(g_j(a)) - p_j(g_j(a))}\right)\right).$$

La matrice de crédibilité est ensuite utilisée dans un algorithme itératif qui sera exécuté deux fois. Une première fois pour trouver la ou les meilleures actions jusqu'à la ou les moins bonnes, on appelle cette étape la distillation descendante, et une deuxième fois pour trouver la ou les actions les plus mauvaises jusqu'à la ou les meilleures, on appelle cette étape la distillation ascendante. L'algorithme consiste à l'itération i à extraire de l'ensemble A un sous-ensemble Ai composé des meilleures ou des plus mauvaises actions selon l'étape de distillation. La règle de sélection est basée sur un seuil de discrimination calculé selon la formule suivante :

$$s_i(\lambda) = \alpha + \lambda_i \times \beta,$$

Où α = 0.30, β = – 0.15 et λ_i représente la plus grande valeur de tous les degrés de crédibilité.

La comparaison des deux distillations fournit la matrice de surclassement finale. Le classement final est lui déduit en fonction du nombre d'occurrences I, R, P et NP de la matrice de surclassement.

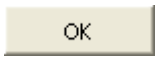
Pour offrir de la souplesse à la méthode, des options sont proposées à l'utilisateur. La première permet de choisir le sens de préférence des performances pour chaque critère. Un sens croissant signifie que les performances vont de la plus petite à la plus grande et que la préférence est donnée aux performances élevées (maximisation du critère). A l'inverse, un sens décroissant signifie que les performances vont de la plus grande à la plus petite et que la préférence est donnée aux valeurs faibles (minimisation du critère). Cette option est codée dans le logiciel par la valeur 1 pour maximiser le critère et la valeur -1 pour minimiser le critère. La valeur 1 est prescrite par défaut. Les deux options suivantes sont dédiées au format et à la direction des seuils. Ces derniers peuvent être définis constant ou comme une fonction linéaire

de la performance gj(a). Le choix se portera sur la fonction linéaire si la différence uj est grande. Dans ce cas, les seuils sont définis soit en mode direct c'est-à-dire que la performance de l'action utilisée dans le calcul des seuils est celle pour laquelle l'action est la plus mauvaise, soit en mode inverse, la performance utilisée sera celle de la meilleure des deux actions. Cette dernière option prend la valeur 1 lorsque le mode est direct, -1 si le mode est inverse. La valeur 1 est prescrite par défaut.

Pour utiliser la méthode Electre 3, l'utilisateur doit fournir en plus des éléments communs aux deux méthodes cités plus haut, les seuils d'indifférence, de préférence et véto et choisir le format des seuils.

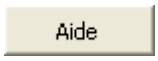
Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.



: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.

Onglet **Général**:

Tableau critères/actions : vous pouvez sélectionner un tableau comprenant les performances des actions qui se compose de n critères en ligne et p actions en colonne. Les données doivent être de type numérique. Si des en-têtes de colonne ont été sélectionnés pour les actions, veuillez vérifier que l'option « Libellés des variables » est activée.

Poids des critères : vous sélectionnez un tableau comprenant le poids de chaque critères. Les données doivent être de type numérique. Ce tableau doit avoir autant de lignes que le tableau des critères/actions et une seule colonne. Si un libellé est sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée. Les valeurs manquantes ne sont pas acceptées.

Choix de la méthode : vous pouvez choisir entre les méthodes Electre 1 ou 3 (voir la section description pour plus de détails sur les méthodes).

Seuil de concordance : Seuil utilisé dans la méthode Electre 1. Vous choisissez la valeur du seuil de concordance entre 0 et 1 (valeur par défaut : 1). Ce seuil doit être plus grand que le seuil de discordance.

Seuil de discordance : Seuil utilisé dans la méthode Electre 1. Vous choisissez la valeur du seuil de discordance entre 0 et 1 (valeur par défaut : 0). Ce seuil doit être plus petit que le seuil de concordance.

Seuil d'indifférence : Seuil utilisé dans la méthode Electre 3. Vous pouvez sélectionner un tableau comprenant le seuil d'indifférence de chaque critères. Les données doivent être de type numérique. Ce tableau doit avoir autant de lignes que le tableau des critères/actions et 1 ou 2 colonnes selon le format choisi (« voir la selection format des seuils »). Si un libellé est sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée. Les valeurs manquantes ne sont pas acceptées pour cette variable.

Seuil de préférence : Seuil utilisé dans la méthode Electre 3. Vous pouvez sélectionner un tableau comprenant le seuil de préférence de chaque critères. Les données doivent être de type numérique. Ce tableau doit avoir autant de lignes que le tableau des critères/actions et 1 ou 2 colonnes selon le format choisi (voir la sélection « format des seuils »). Si un libellé est sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée. Les valeurs manquantes ne sont pas acceptées pour cette variable.

Seuil véto : Seuil utilisé dans la méthode Electre 3. Vous pouvez sélectionner un tableau comprenant le seuil véto de chaque critères. Les données doivent être de type numérique. Ce tableau doit avoir autant de lignes que le tableau des critères/actions et 1 ou 2 colonnes selon le format choisi (voir la sélection « format des seuils »). Si un libellé est sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée. Les valeurs manquantes ne sont pas acceptées pour cette variable.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des variables : activez cette option si la première ligne des données sélectionnées (variables dépendantes et explicatives, libellés des observations, poids) contient un libellé.

Libellés des critères : vous pouvez activer cette option si vous souhaitez sélectionner un tableau contenant les libellés de chaque critère. Les données doivent être de type caractère. Par défaut les libellés sont notés « Crit.1, Crit.2, … ». Ce tableau doit avoir autant de lignes que le tableau des critères/actions. Si le libellé des variables a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

Format des seuils : Utilisé dans la méthode Electre 3. Vous pouvez choisir le format des seuils entre constant ou linéaire. Dans le cas constant, une seule valeur est attendue par seuil et par critère. Dans le cas linéaire, deux valeurs sont attendues sur deux colonnes correspondant à la pente (première colonne) et à l'interception (deuxième colonne).

Onglet **Options** :

Direction d'évaluation : activez cette option pour ajouter une condition sur le sens d'évaluation de chaque critère. Vous devez sélectionner un tableau comprenant autant de lignes que le tableau des critères/actions et une colonne. Les données doivent être de type numérique égales à 1 ou -1. Par défaut la valeur est fixée à 1. Si un libellé est sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée. Les valeurs manquantes ne sont pas acceptées pour cette variable.

Direction des seuils : activez cette option pour ajouter une condition sur le sens des seuils pour chaque critère. Vous devez sélectionner un tableau comprenant autant de lignes que le tableau des critères/actions et 1 colonne. Les données doivent être de type numérique égales à 1 ou -1. Par défaut la valeur est fixée à 1. Si un libellé est sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée. Les valeurs manquantes ne sont pas acceptées pour cette variable.

Onglet **Données manquantes** :

Ne pas accepter les valeurs manquantes : activez cette option pour que XLSTAT empêche la poursuite des calculs si des valeurs manquantes sont détectées.

Supprimer les critères : activez cette option pour supprimer les critères (lignes) comportant des données manquantes dans le tableau des critères/actions.

Supprimer les actions : activez cette option pour supprimer les actions comportant des données manquantes.

Estimer les données manquantes : activez cette option pour estimer les données manquantes avant le début des calculs.

- **Moyenne** : activez cette option pour estimer les données manquantes en utilisant la moyenne (variables quantitatives) des actions.
- **Plus proche voisin** : activez cette option pour estimer les données manquantes d'un critère en recherchant le plus proche voisin de celui-ci.

Onglet **Sorties** :

Statistiques descriptives : activez cette option pour afficher les statistiques descriptives. Les tableaux de statistiques descriptives présentent pour toutes les actions sélectionnées des statistiques simples.

Matrice de concordance : activez cette option pour afficher la matrice de concordance. Pour Electre 3, c'est la matrice globale qui est affichée.

Matrice de discordance : activez cette option pour afficher la matrice de discordance. Cette option est disponible pour la méthode Electre 1.

Matrice de crédibilité : activez cette option pour afficher la matrice de crédibilité. Cette option est disponible pour la méthode Electre 3.

Matrice de surclassement : activez cette option pour afficher la matrice de surclassement.

Tableau de classement : activez cette option pour afficher le tableau de classement.

Analyse de sensibilité :Cette option est disponible en activant la sortie « Tableau de classement » dans la méthode Electre 1. Elle permet d'afficher les tableaux de classement obtenus en opérant une analyse de sensibilité de 10% sur les seuils de concordance et de discordance.

Résultats

XLSTAT propose un nombre important de tableaux et de graphiques afin de faciliter l'analyse et l'interprétation des résultats.

Statistiques descriptives : le résultat, sous forme de tableau, affiche le nombre d'observations, le nombre d'observations manquantes, le minimum, le maximum, la moyenne, et l'écart-type non biaisé pour chaque action.

Matrice de concordance : le résultat, sous forme de tableau, affiche les indices de la matrice de concordance calculés selon l'équation (1) pour la méthode Electre 1 ou selon l'équation (5) pour la méthode Electre 3 (voir la section description).

Matrice de discordance : le résultat, sous forme de tableau, affiche les indices de la matrice de discordance calculés selon l'équation (2) pour la méthode Electre 1 (voir la section description).

Matrice de crédibilité : le résultat, sous forme de tableau, affiche les indices de la matrice de crédibilité calculés selon l'équation (4) donnée dans la section description de la méthode Electre 3.

Matrice de surclassement : le résultat, sous forme de tableau, affiche les 0 et les 1 obtenus avec les relations de surclassement (3) pour la méthode Electre 1. Pour la méthode Electre 3, il affiche les caractères I, R, P et NP pour respectivement indifférent, incomparable, préféré et non préféré.

Tableau de classement : dans ce tableau est affiché le classement final des actions.

Tableaux de classement après analyse de sensibilité sur le seuil de concordance : dans ces tableaux sont affichés les classements finaux des actions obtenus avec la méthode Electre 1 our un seuil de concordance modifié et le seuil de discordance donné par l'utilisateur (sinon égale à sa valeur par défaut 0). Le tableau de gauche est le résultat avec une augmentation de la valeur donnée de 10%, le tableau de droite est le résultat avec une diminution de 10%.

Tableaux de classement après analyse de sensibilité sur le seuil de dis cordance : dans ces tableaux sont affichés les classements finaux des actions obtenus avec la méthode Electre 1 pour un seuil de concordance donné par l'utilisateur (sinon égale à sa valeur par défaut 1) et le seuil de discordance modifié. Le tableau de gauche est le résultat avec une augmentation de la valeur donnée de 10%, le tableau de droite est le résultat avec une diminution de 10%.

Exemple

Un exemple d'utilisation de la méthode Electre 1 est disponible sur le Centre d'aide XLSTAT à l'adresse :

http://www.xlstat.com/demo-elcf.htm

Un exemple d'utilisation de la méthode Electre 3 est disponible sur le Centre d'aide XLSTAT à l'adresse :

http://www.xlstat.com/demo-elc3f.htm

Bibliographie

Bouyssou, D., Roy, B. (1986). La notion de seuils de discrimination en analyse multicritère. Information Systems and Operational Research. Vol. 25, n°4,1987.

Nafi, A., and Werey, C. (2009). Aide à la décision multicritère : introduction aux méthodes d'analyse multicritère de type ELECTRE. Notes de cours, Module « Ingénierie financière », ENGEES.

Valléé, D., Zielniewicz, P., Roy, B. (1994). ELECTRE III-IV, version 3.X. Aspects méthodologiques (tome 1). La collection des cahiers et Documents du LAMSADE.

Vetschera, R. (1986). Sensitivity Analysis for the ELECTRE Multicriteria Method. Zeitschrift Operations Research. Vol. 30, 99-117.

Roy, B. (1977). Electre III, un algorithme de classement fondé sur une représentation floue des préférences en présence de critères multiples. Cahiers du Centre d'études de recherche opérationnelle, 20 (1) : 3-24

Plans d'expériences pour l'analyse hiérarchique des procédés

Utilisez cet outil pour générer les plans d'expériences nécessaires à une Analyse Hiérarchique des Procédés (AHP).

Dans cette section :

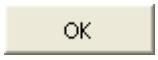
Description

Le principe de l'Analyse Hiérarchique des Procédés (AHP) est de comparer les solutions d'un problème décisionnel sur un ensemble de critères afin d'en déduire la meilleure solution. L'application d'une telle méthode nécessite de définir une multitude de tableaux de comparaison (ou matrices de comparaison) dont le nombre peut vite augmenter selon les caractéristiques du problème décisionnel (nombre de critères, sous-critères, d'alternatives et d'évaluateurs, voir l'aide de la méthode AHP). Dans un cas simple de 4 alternatives, 4 critères et 2 évaluateurs, 9 tableaux doivent être définis et remplis pour permettre les calculs de la méthode AHP.

Pour limiter les erreurs de saisi XLSTAT propose l'outil DHP pour générer automatiquement les tableaux de comparaison utiles à une analyse AHP. Son utilisation se fait simplement en précisant les caractéristiques du problème décisionnel, soit à minima la liste des alternatives, la liste des critères et des sous-critères s'il en existe. La liste des évaluateurs peut aussi être ajoutée en option (voir description de la méthode AHP pour la définition des termes).

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.

 : cliquez sur ce bouton pour rétablir les options par défaut.

 : cliquez sur ce bouton pour effacer les sélections de données.

 : cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.

Onglet **Général**:

Alternatives : sélectionnez un tableau comprenant la liste des alternatives. Les données doivent être des chaines de caractères. Si des en-têtes de colonne ont été sélectionnés pour les alternatives, veuillez vérifier que l'option « Libellés des variables » est activée. Les données manquantes ne sont pas acceptées.

Critères : sélectionnez un tableau comprenant la liste des critères. Les données doivent être des chaines de caractères. Si un libellé est sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée. Les données manquantes ne sont pas acceptées.

Sous-critères : activez cette option si vous souhaitez sélectionner un tableau contenant la liste des critères. Les données doivent être des chaines de caractères. Ce tableau doit avoir autant de colonnes que le nombre de critères sélectionné. Si un libellé est sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée. Les données manquantes ne sont pas acceptées.

Nombre d'évaluateurs : saisissez le nombre d'évaluateurs.

Libellés des évaluateurs : activez cette option si vous souhaitez sélectionner un tableau contenant la liste des libellés de chaque évaluateur. Les données doivent être des chaines de caractères. Si l'option n'est pas cochée, le libellé est de type numérique, soit 1, 2, 3, etc.... Si un en-tête est sélectionné au tableau, veuillez vérifier que l'option « Libellés des variables » est activée. Les données manquantes ne sont pas acceptées.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des variables :activez cette option si la première ligne des données sélectionnées (alternatives, critères, sous-critères, libellés des évaluateurs) contient un libellé.

Résultats

Tableau récapitulatif : ce tableau regroupe l'ensemble des données sélectionnées en affichant dans cet ordre la liste des critères, des sous-critères si l'option est cochée, des alternatives et des évaluateurs si l'option est cochée.

Tableau de Saaty : ce tableau contient les valeurs de Saaty, une définition de chaque valeur et des commentaires. En dessous du tableau quelques phrases expliquent comment utiliser les valeurs de Saaty.

Matrices de comparaison de l'évaluateur X : sont les tableaux de comparaison que doit remplir l'évaluateur x. Ils sont affichés sur 2 ou 3 lignes selon le niveau de hiérarchisation (voir méthode AHP). Sur la première ligne est affichée la matrice de comparaison des critères. Sur la deuxième ligne sont affichées les matrices de comparaison des sous-critères si l'option « sous-critère » est sélectionnée. Sur la troisième ligne de tableau sont affichées les matrices de comparaison des alternatives si l'option « sous-critères » est cochée, sinon elles sont affichées sur la deuxième ligne de tableau.

Matrice de comparaison sur les critères : est le tableau de comparaison des critères. Il est composé de 1 sur la diagonale et de cellules vides. La partie en dessous de la diagonale est grisée. Seules les cellules au-dessus de la diagonale peuvent être saisies (voir méthode AHP).

Matrice de comparaison sur les sous-critères : est le tableau de comparaison des sous-critères. Si plusieurs critères ont des sous-critères, les tableaux sont alignés. Chaque matrice est composée de 1 sur la diagonale et de cellules vides. La partie en dessous de la diagonale est grisée. Seules les cellules au-dessus de la diagonale peuvent être saisies (voir méthode AHP).

Matrices de comparaison des alternatives : sont les tableaux de comparaison des alternatives. Ils sont affichés sur une même ligne en fonction des critères et des sous-critères donnés. Les matrices sont composées de 1 sur la diagonale et de cellules vides. La partie en dessous de la diagonale est grisée. Seules les cellules au-dessus de la diagonale peuvent être saisies (voir méthode AHP).

Exemple

Un exemple d'utilisation de la méthode AHP est disponible sur le Centre d'aide XLSTAT à l'adresse :

http://www.xlstat.com/demo-ahpf.htm

Aide Multicritère à la décision : méthode AHP

Utilisez cette application pour résoudre votre problème décisionnel par une Analyse Hiérarchique des Procédés (AHP). Elle doit être utilisée sur un plan d'expérience, qui doit être généré par l'outil DHP d'XLSTAT disponible également dans les applications d'aide à la décision.

La méthode AHP est une méthode d'aide à la décision se basant sur la hiérarchisation des critères. Elle sera privilégiée lorsque le nombre de critères reste raisonnable, et lorsque l'utilisateur est capable d'évaluer 2 à 2 les éléments de son problème. La version d'AHP proposée dans XLSTAT à l'avantage de ne pas avoir de limitations sur le nombre de critères, de sous-critères et d'alternatives et autorise la participation d'un grand nombre d'évaluateurs.

Dans cette section :

Description

La méthode AHP est une méthode adaptée aux problèmes de décision multicritères c'est-à-dire comportant plusieurs solutions satisfaisant un ensemble de critères. L'approche de la méthode consiste à simplifier le problème en le décomposant en un système hiérarchique. Thomas Saaty est à l'origine de cette méthode et l'a créé dans les années 1970.

On appelle *alternatives* les solutions du problème décisionnel, *critères* les paramètres sur lesquels sont évaluées les alternatives, *sous-critères* les paramètres appartenant à un critère et sur lesquels sont évaluées les alternatives et *évaluateur* la personne qui va faire les évaluations. On parle d'un problème à 2 niveaux lorsqu'il admet des sous-critères, à l'inverse c'est un problème de niveau 1.

Le principe de la méthode repose sur l'évaluation 2 à 2 des éléments du problème qui est regroupée dans des tableaux de comparaison. Ils sont définis à chaque niveau de la hiérarchisation. Au niveau 0 l'utilisateur définit le tableau de comparaison des critères, au niveau 1 ceux des sous-critères si c'est un problème à 2 niveaux sinon les tableaux de comparaison des alternatives sur les critères. Enfin, au niveau 2 on définit les tableaux de comparaison des alternatives sur les critères et/ou les sous-critères. L'ensemble de tous ces tableaux forme le plan d'expérience pour une analyse AHP.

XLSTAT met à votre disposition l'outil DHP pour créer votre plan d'expérience. Il est disponible dans les applications « aide à la décision ». Vous trouverez l'aide d'utilisation ici.

Les tableaux de comparaison doivent être complétés par l'utilisateur en fonction de valeurs choisies dans le tableau de Saaty reporté ci-dessous. Saaty a défini une échelle d'évaluation qui mesure l'importance ou la différence d'un élément sur un autre.

Appréciation	Degré d'importance
Importance égale de deux éléments	1
Importance modérée d'un élément par rapport à un autre	3
Importance forte d'un élément par rapport à un autre	5
Importance très forte d'un élément par rapport à un autre	7
Importance extrême d'un élément par rapport à un autre	9
Exprime des valeurs intermédiaires	2, 4, 6, 8
Réciprocité	1/degré d'importance

Les premiers calculs de la méthode AHP vont porter sur le calcul du vecteur des priorités à partir des valeurs du tableau de comparaison, c'est à dire le poids de chaque critère. La formule appliquée est la suivante :

```
Poids du critère = somme des lignes normalisées / nombre de critère
```

Avec la même formule, le vecteur des priorités des sous-critères est calculé pour chaque critère. Ce vecteur est ensuite pondéré par le poids du critère. On obtient ainsi les poids de chaque critère et des sous-critères qui interviendront dans la pondération des vecteurs priorités des alternatives calculées avec la même formule mathématique.

Une option dans les résultats de sortie est proposée pour évaluer la cohérence des données. Ce test permet de contrôler la saisie des valeurs dans les tableaux de comparaison. En effet si l'alternative A1 est évaluée 2 fois plus grande que l'alternative A2, et A2 est jugée 3 fois plus grande que l'alternative A3 et A3 4 fois plus grande que A1, alors le test permettra de dire qu'il y a une incohérence dans les données. On la mesure avec 2 paramètres : l'indice de cohérence (IC) et le ratio de cohérence (RC). La formule de calcul pour l'indice de cohérence est la suivante.

```
IC = cohérence moyenne - nombre d'éléments / (nombre d'éléments -1)
```

Le nombre d'éléments est le nombre de colonne ou de ligne du tableau de comparaison. Pour obtenir la cohérence moyenne on multiplie d'abord la matrice de comparaison avec son vecteur des priorités qui nous donne ainsi un nouveau vecteur. Puis, on divise ce dernier par le poids du vecteur priorité de l'élément de la même ligne. La moyenne de ce vecteur normé donne la cohérence moyenne. La formule pour le calcul du ratio de cohérence est donnée par :

```
RC = Indice de cohérence / cohérence aléatoire
```

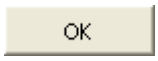
Où la cohérence aléatoire est donnée par le tableau suivant :

Nombre de critères	2	3	4	5	6	7	8	9
Cohérence aléatoire	0.00	0.58	0.90	1.12	1.24	1.32	1.41	1.45

Si le ratio de cohérence est inférieur ou égal à 10% alors l'évaluation est jugée cohérente. A l'inverse, s'il est plus grand que 10% il est recommandé de revoir l'évaluation du tableau de comparaison concerné.

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.

Onglet **Général** :

Feuille du plan d'expérience : sélectionnez une feuille contenant le plan d'expérience DHP généré par XLSTAT (voir l'aide de la méthode DHP) que vous souhaitez analyser.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des variables : activez cette option si la première ligne des données sélectionnées (alternatives, critères, sous-critères, libellés des évaluateurs) contient un en-tête.

Onglet **Données manquantes** :

Ne pas accepter les valeurs manquantes : activez cette option pour que XLSTAT empêche la poursuite des calculs si des valeurs manquantes sont détectées.

Remplacer avec la valeur d'importance égale 1 : activez cette option pour remplacer les données manquantes dans un tableau de comparaison par la valeur d'importance 1 du tableau

de Saaty (voir section description). Ceci donnera une importance égale entre les 2 éléments de comparaison.

Estimer les données manquantes : activez cette option pour estimer les données manquantes avant le début des calculs.

- **Moyenne** : activez cette option pour estimer les données manquantes en utilisant la moyenne (variables quantitatives) des actions.
- **Plus proche voisin** : activez cette option pour estimer les données manquantes d'un critère en recherchant le plus proche voisin de celui-ci.

Onglet **Sorties** :

Tableau moyen des résultats : activez cette option pour afficher le ou les tableaux moyens des résultats obtenus par évaluateur. Cette option est valable à partir de 2 évaluateurs.

Tableau moyen des critères : activez cette option pour afficher le tableau moyen des résultats sur les critères obtenus par évaluateur. Cette option est valable à partir de 2 évaluateurs.

Tableau moyen des sous-critères : activez cette option pour afficher le tableau moyen des résultats sur les sous-critères obtenus par évaluateur. Cette option est valable à partir de 2 évaluateurs et lorsque des sous-critères sont sélectionnés.

Tableau moyen des alternatives : activez cette option pour afficher le tableau moyen des résultats sur les alternatives obtenues par évaluateur. Cette option est valable à partir de 2 évaluateurs.

Résultats par évaluateur : activez cette option pour afficher le ou les tableaux de résultats obtenus pour chaque évaluateur.

Critères : activez cette option pour afficher les tableaux de résultats sur les critères obtenus pour chaque évaluateur.

Sous-critères : activez cette option pour afficher les tableaux de résultats sur les sous-critères obtenus pour chaque évaluateur.

Alternatives : activez cette option pour afficher les tableaux de résultats sur les alternatives obtenus pour chaque évaluateur.

Consistance des données : activez cette option pour afficher le calcul de consistance des données. Cette option est disponible en activant l'option « Résultats par évaluateur ».

Index de consistance : activez cette option pour afficher le calcul d'index de consistance des données (voir description pour plus de détail). Cette option est disponible en activant l'option « Résultats par évaluateur ».

Ratio de consistance : activez cette option pour afficher le calcul de ratio de consistance des données (voir description pour plus de détail). Cette option est disponible en activant l'option « Résultats par évaluateur ».

Onglet **Graphiques** :

Diagrammes en bâton : activez cette option pour afficher le ou les tableaux de résultats obtenus par évaluateur sous forme de diagramme en bâton. Cette option est disponible que si des tableaux de résultats sont demandés.

Critères : activez cette option pour afficher les tableaux de résultats obtenus sur les critères par évaluateur sous forme de diagramme en bâton. Cette option est disponible que si l'option « Diagrammes en bâton » est activée et si les tableaux de résultats sur les critères sont demandés.

Sous-Critères : activez cette option pour afficher les tableaux de résultats obtenus sur les sous-critères par évaluateur sous forme de diagrammes en bâton. Cette option est disponible que si l'option « Diagrammes en bâton » est activée et si les tableaux de résultats sur les sous-critères sont demandés.

Alternatives : activez cette option pour afficher les tableaux de résultats obtenus sur les alternatives par évaluateur sous forme de diagrammes en bâton. Cette option est disponible que si l'option « Diagrammes en bâton » est activée et si les tableaux de résultats sur les alternatives sont demandés.

Résultats

Priorités moyennes par critère : ce résultat, sous forme de tableau, correspond aux pourcentages relatifs moyens des vecteurs poids par critère sur l'ensemble des évaluateurs. Si les options « Diagrammes en bâton » et « critères » sont sélectionnées le résultat est aussi affiché sous forme de diagrammes en bâton en dessous du tableau.

Priorités moyennes par sous-critère : ce résultat, sous forme de tableau, correspond aux pourcentages relatifs moyens des vecteurs poids par sous-critère sur l'ensemble des évaluateurs. Si les options « Diagrammes en bâton » et « sous-critères » sont sélectionnées le résultat est aussi affiché sous forme de diagrammes en bâton en dessous des tableaux.

Priorités moyennes par alternatives : ce résultat, sous forme de tableau, correspond aux pourcentages relatifs moyens des vecteurs poids par alternatives sur l'ensembles des évaluateurs. Si l'option « Diagrammes en bâton » et « alternatives » sont sélectionnées le résultat est aussi affiché sous forme de diagrammes en bâton en dessous du tableau.

Résultats obtenus à partir des notes de l'évaluateur x : donnent l'ensemble des tableaux et des diagrammes en bâton pour l'évaluateur x en fonction des options cochées :

- **Les priorités par critère** : sont les pourcentages relatifs pour chaque critère.
- **Les priorités par sous-critère du critère XXXX** : sont les pourcentages relatifs pour chaque sous-critère du critère XXXX.
- **Les priorités par alternative** : sont les pourcentages relatifs pour chaque alternative.

Si l'option « Consistance des données » est activée les variables IC et RC sont calculées et affichées en dessous des tableaux (voir la section description pour plus de détails).

Exemple

Un exemple d'utilisation de la méthode AHP est disponible sur le Centre d'aide XLSTAT à l'adresse :

http://www.xlstat.com/demo-ahpf.htm

Bibliographie

Saaty, T. (1977). A scaling method for priorities in Hierarchical Structures. Journal of mathematical psychology, Vol. 15, 234-281.

Saaty, T. (1978). Exploring the interface between hierarchies, multiple objectives and fuzzy set. Fuzzy sets and systems, Vol. 1, 57-68.

Arbres de décision

La fonctionnalité Arbre de décision de XLSTAT est un outil d'aide à la décision dont le résultat final représente un ensemble de choix sous la forme graphique d'un arbre. Les différentes décisions possibles sont représentées par des nœuds et situées aux extrémités de branches (les « feuilles » de l'arbre). Elles sont atteintes en fonction de décisions prises à chaque étape. Largement utilisé dans divers domaines, sa fonction principale est de déterminer un chemin optimal selon des critères paramétrables.

Dans cette section :

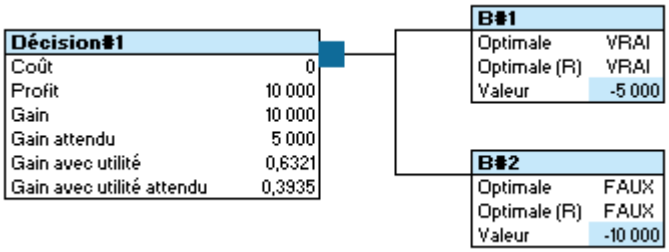
Description

Un arbre de décision est un schéma représenté par un ensemble de nœuds interconnectés car reliés par des branches. Il permet à son créateur d'évaluer différentes actions possibles en fonction de leur coût, leur bénéfice et leur probabilité. Il commence généralement par un nœud duquel découlent plusieurs résultats possibles. Chacun de ces résultats mène à d'autres nœuds, les enfants, desquels émanent d'autres possibilités. Le schéma ainsi obtenu rappelle la forme d'un arbre.

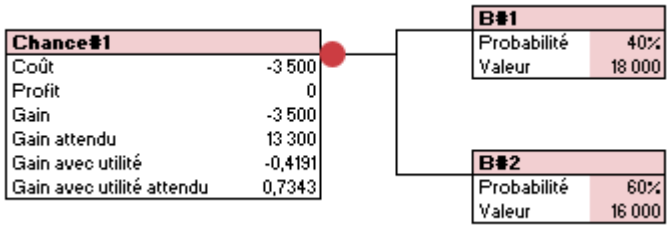
Différents types de nœuds

Voici une brève description pour chaque type de nœud, avec leur représentation graphique, sous forme de bloc, dans XLSTAT :

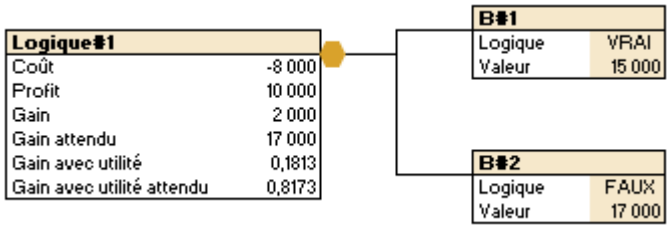
- **Nœud décision** : représenté par un carré bleu, il illustre une décision à prendre parmi plusieurs choix (branches) possibles. Voici un exemple de nœud décision.



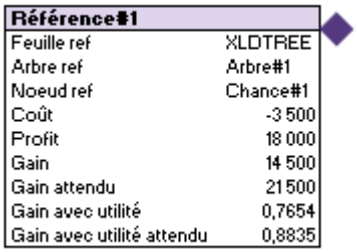
- **Nœud chance** : représenté par un cercle rouge, il offre différents résultats, chacun ayant un % de chance de réalisation. La somme de ces pourcentages doit être de 100. Voici un exemple de nœud chance.



- **Nœud logique** : repésenté par un hexagone jaune, il impose un résultat ou non en fonction du retour d'une formule logique. Voici un exemple de nœud logique.



- **Nœud référence** : représenté par un losange violet, il fait référence à un sous-arbre, c'est-à-dire, à un autre nœud avec tous ses enfants. Voici un exemple de nœud référence.



- **Nœud final** : représenté par un triangle vert, il est le résultat final d'un chemin de décision. Voici un exemple de nœud final.

Informations disponibles pour chaque nœud

Comme vous pouvez le constater sur les figures précédentes, chaque nœud comprend différentes informations qui sont à votre disposition. Vous pouvez choisir d'afficher toutes ces informations ou non. Ce paramétrage sera abordé dans la partie Construction d'un arbre. En attendant, voici une description de ces différentes informations. Les notions de mode de calcul et fonction d'utilité seront présentées ensuite.

- **Coût** : le coût d'un nœud est la somme de tous les coûts de ses nœuds parents.
- **Profit** : le profit d'un nœud est la somme de tous les profits de ses nœuds parents.
- **Gain** : le gain d'un nœud est la somme de son coût et de son profit. Il est donc calculé en fonction des coûts et profits de ses nœuds parents. Le coût étant bien entendu une valeur négative et le profit une valeur positive.
- **Gain attendu** : le gain attendu d'un nœud dépend du gain de ses nœuds enfants. Il va permettre de prendre une décision en fonction du mode de calcul choisi.
- **Gain avec utilité** : le gain avec utilité d'un nœud est calculé grâce à la fonction d'utilité exponentielle appliquée au gain de ce même nœud.
- **Gain avec utilité attendu** : le gain avec utilité attendu d'un nœud dépend du gain avec utilité de ses nœuds enfants. Il va permettre de prendre une décision en fonction du mode de calcul choisi.
- **Feuille référente** : seulement dans le cas d'un nœud référence, c'est le nom de la feuille où se trouve le nœud référent. Si cette feuille est dans un autre classeur que celui où se trouve l'arbre courant alors le nom de ce classeur est également renseigné.
- **Arbre référent** : seulement dans le cas d'un nœud référence, c'est le nom de l'arbre contenant le nœud référent.
- **Nœud référent** : seulement dans le cas d'un nœud référence, c'est le nom du nœud référent.

Modes de calculs

Il existe deux modes de calcul pour vous aider dans la prise de décision. Vous pouvez choisir de maximiser votre gain si vous souhaitez optimiser votre profit ou alors de minimiser votre gain si vous souhaitez optimiser votre coût. Ces deux modes de calcul sont agrémentés de la possibilité d'avoir un gain, somme des coûts et profits, mais aussi un gain calculé à partir d'une fonction d'utilité exponentielle.

Dans un univers risqué, un individu parfaitement rationnel prend des décisions d'investissement en maximisant l'espérance (au sens probabiliste du terme) de sa fonction d'utilité. Celle-ci traduit la satisfaction engendrée par une richesse future donnée. Dans notre cas, nous utilisons la fonction d'utilité exponentielle définie comme suit :

$$U(x) = \frac{1 - \exp(-Rx)}{R}, R \neq 0$$

$$U(x) = x, R = 0$$

où x représente le gain et R l'utilité ou degré d'aversion au risque avec $R > 0$ en cas d'aversion au risque, $R = 0$ en cas de neutralité vis-à-vis du risque et $R < 0$ si le risque est recherché. Il est facile de constater qu'une utilité $R = 0$ revient à avoir un gain avec utilité équivalent à un gain sans utilité.

Voici une description des calculs effectués selon le type de nœud. Nous commençons par le nœud final car le gain, avec utilité ou non, attendu d'un nœud dépend de cette même information pour un nœud enfant.

- **Nœud final** : dans le cas du nœud final, le gain attendu correspond au gain et le gain avec utilité attendu correspond au gain avec utilité.
- **Nœud décision** : le gain, avec utilité ou non, attendu dépend du mode de calcul.
    - **Maximiser le gain sans fonction d'utilité** : le gain attendu correspond au plus grand gain attendu parmi les gains attendus des nœuds enfants directs.
    - **Maximiser le gain avec fonction d'utilité** : le gain avec utilité attendu correspond au plus grand gain avec utilité attendu parmi les gains avec utilité attendus des nœuds enfants directs.
    - **Minimiser le gain sans fonction d'utilité** : le gain attendu correspond au plus petit gain attendu parmi les gains attendus des nœuds enfants directs.
    - **Minimiser le gain avec fonction d'utilité** : le gain avec utilité attendu correspond au plus grand gain avec utilité attendu parmi les gains avec utilité attendus des nœuds enfants directs. Dans ce cas, la fonction d'utilité devient :

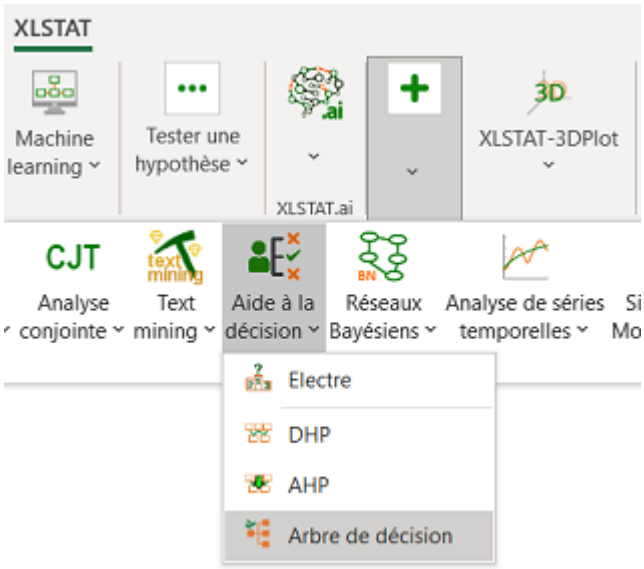
$$U(x) = \frac{1 - \exp(Rx)}{R}, R \neq 0$$

$$U(x) = x, R = 0$$

- **Nœud chance** : le gain, avec utilité ou non, attendu ne dépend pas du mode de calcul. Pour chaque nœud enfant direct, il suffit de faire la multiplication de son gain attendu par sa probabilité de réalisation. Le gain attendu du nœud parent est ainsi la somme de ces multiplications. L'opération est la même dans le cas du gain avec utilité attendu.
- **Nœud logique** : le gain attendu, avec utilité ou non, ne dépend pas du mode de calcul. Il est celui de l'unique nœud enfant dont la formule logique est vérifiée. Si aucun nœud enfant n'a sa formule logique vérifiée alors aucun gain attendu, avec utilité ou non, ne peut être calculé. En aucun cas ce nœud ne fera partie des décisions possibles.
- **Nœud référence** : un nœud référence a pour référent un nœud décision, chance ou logique. Il suffit donc de se reporter à ces types de nœuds pour connaître le calcul effectué.

Barre d'outils

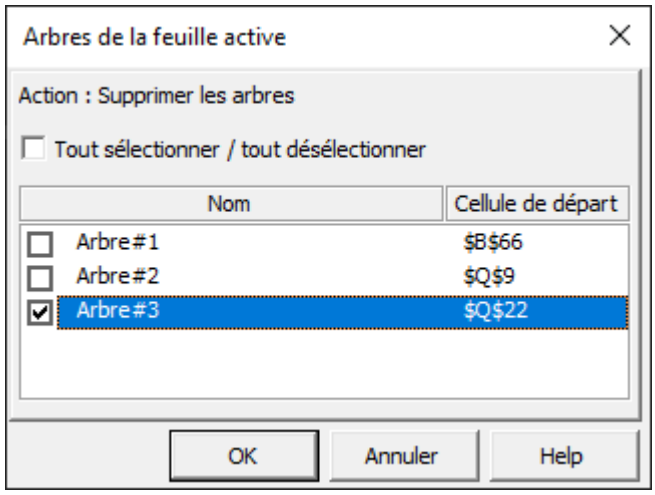
Pour construire un nouvel arbre, il faut commencer par afficher le menu relatif aux arbres de décision. Pour cela, lancez XLSTAT et allez chercher l'outil Arbre de décision dans le ruban comme montré dans la figure ci-dessous :



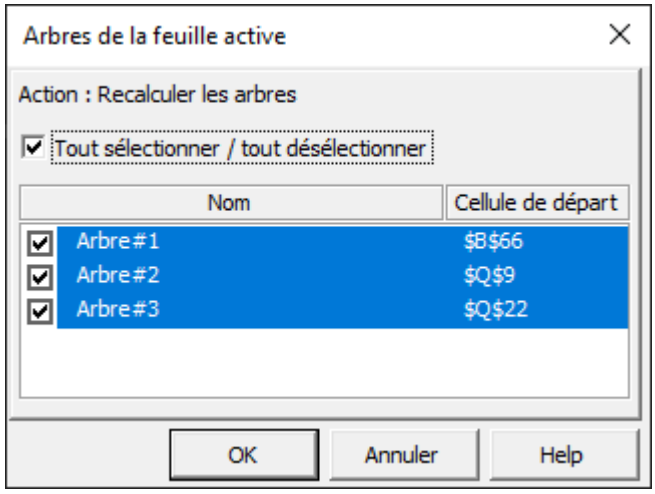
Le menu va ainsi s'afficher :

Voici le détail des différents boutons à votre disposition :

- **Créer un nouvel arbre** : cliquez sur ce bouton si vous souhaitez afficher la boîte de dialogue permettant de créer un nouvel arbre.
- **Supprimer les arbres** : cliquez sur ce bouton si vous souhaitez supprimer des arbres de la feuille active. Une boîte de dialogue s'ouvre alors, afin que vous puissiez sélectionner les arbres à supprimer.



- **Recalculer les arbres** : cliquez sur ce bouton si vous souhaitez recalculer des arbres de la feuille active. Une boîte de dialogue est affichée afin que vous puissiez sélectionner les arbres à recalculer. Cet outil est utile lorsque vous avez plusieurs arbres dans différents onglets. Il peut alors arriver, lorque vous modifiez l'onglet actif, que le résultat des formules des arbres présents dans celui-ci ne soient pas à jour.

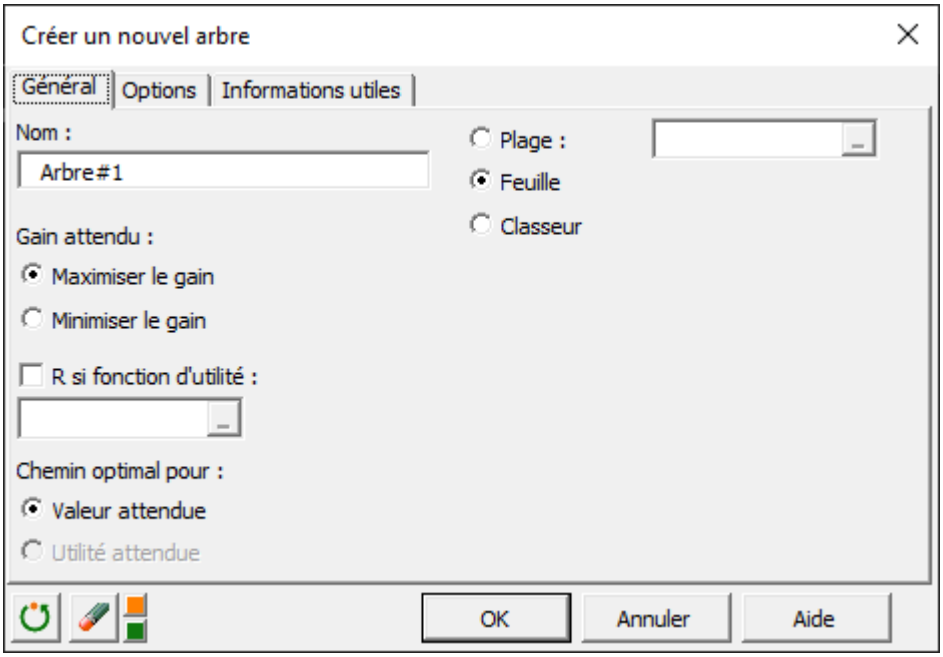


- **Supprimer le quadrillage** : cliquez sur ce bouton si vous souhaitez supprimer le quadrillage de la feuille active (présent seulement si le quadrillage est affiché).
- **Afficher le quadrillage** : cliquez sur ce bouton si vous souhaitez afficher le quadrillage de la feuille active (présent seulement si le quadrillage n'est pas affiché).

Construction d'un arbre

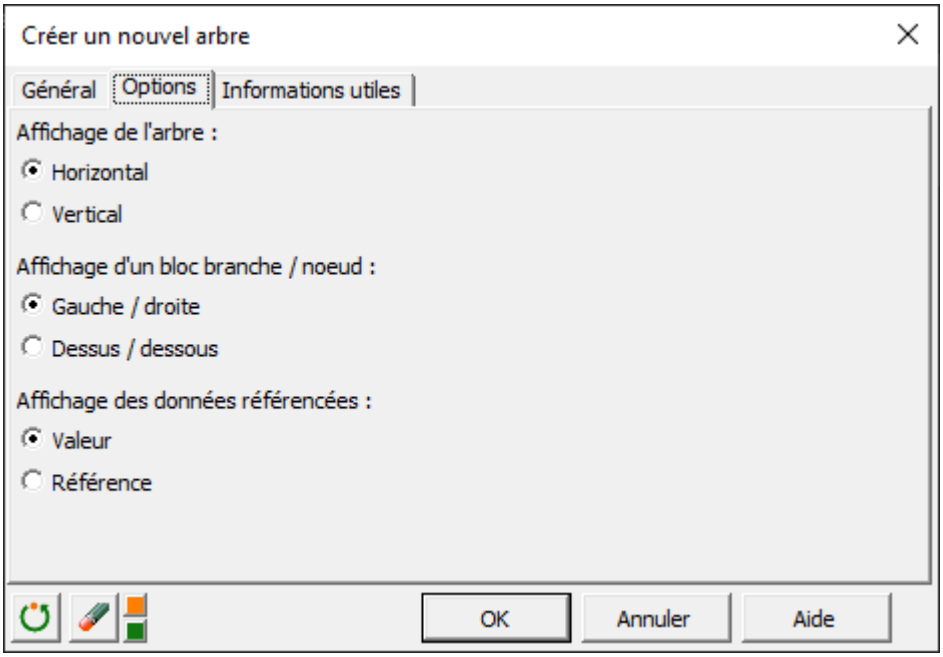
Pour construire un nouvel arbre, il suffit de cliquer sur le bouton correspondant de la barre d'outils présentée précédemment. Une boîte de dialogue s'ouvre alors. Elle permet de paramétrage d'un arbre. Nous détaillons ci-dessous chacun des paramètres proposés. Cette boîte de dialogue peut être ouverte et vos modifiés à tout moment. L'arbre est mis à jour automatiquement si vous modifiez des options.

Onglet **Général** :



- **Nom** : renseignez le nom de l'arbre.
- **Plage** : sélectionnez cette option si vous souhaitez sélectionner vous-même la cellule de départ de l'arbre. Au fur et à mesure de la construction de celui-ci, cette cellule correspond à la cellule en haut à gauche du plus petit bloc contenant l'ensemble de l'arbre.
- **Feuille** : sélectionnez cette option si vous souhaitez que l'arbre se crée dans une nouvelle feuille du classeur actif. La barre d'outils sera automatiquement ajoutée à cette nouvelle feuille.
- **Classeur** : sélectionnez cette option si vous souhaitez que l'arbre se crée dans une feuille d'un nouveau classeur. La barre d'outils sera automatiquement ajoutée à cette feuille.
- **Gain attendu** : sélectionnez ici votre mode de calcul, selon que vous souhaitez **Maximiser le gain** ou **Minimiser le gain**.
- **R si fonction d'utilité** : renseignez ici votre R ou degré d'aversion au risque. Il est également possible de référencer une cellule dans une feuille dont la valeur correspond à votre R.
- **Chemin optimal pour** : sélectionnez le type de chemin optimal que vous souhaitez afficher. Si l'option **Valeur attendue** est sélectionnée alors le chemin optimal sera calculé à partir des gains attendus calculés. Si l'option **Utilité attendue** est sélectionnée alors le chemin optimal sera calculé à partir des gains avec utilité attendus calculés. Cette option ne peut être sélectionnée que si **R si fonction d'utilité** est sélectionnée aussi. *Ce paramétrage n'est pas disponible sous Mac pour l'instant*.

Onglet **Options** :

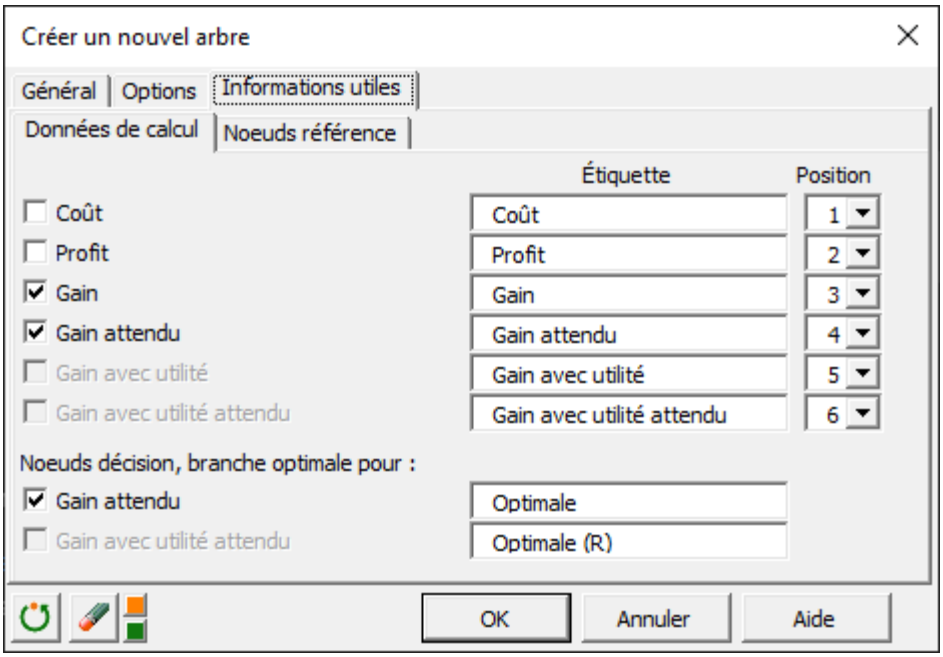


- **Affichage de l'arbre** : sélectionnez le mode d'affichage de l'arbre. Un arbre **horizontal** se construit et progresse de la gauche vers la droite. Un arbre **vertical** se construit et progresse de haut en bas.

- **Affichage d'un bloc branche / nœud** : sélectionnez le mode d'affichage d'un bloc branche par rapport à un bloc nœud. La branche étant la branche parente du bloc nœud. Un affichage **Gauche / droite** va afficher le bloc nœud à droite du bloc branche. Un affichage **Dessus / dessous** va afficher le bloc nœud en dessous du bloc branche.

- **Affichage des données référencées** : sélectionnez le mode d'affichage par défaut des données référencées. Votre choix est appliqué dans l'interface de paramétrage d'un nœud, lorsqu'une donnée peut faire référence à une cellule dans une feuille. Vous choisissez ici le mode d'affichage par défaut mais celui-ci est encore modifiable dans l'interface de paramétrage d'un nœud.

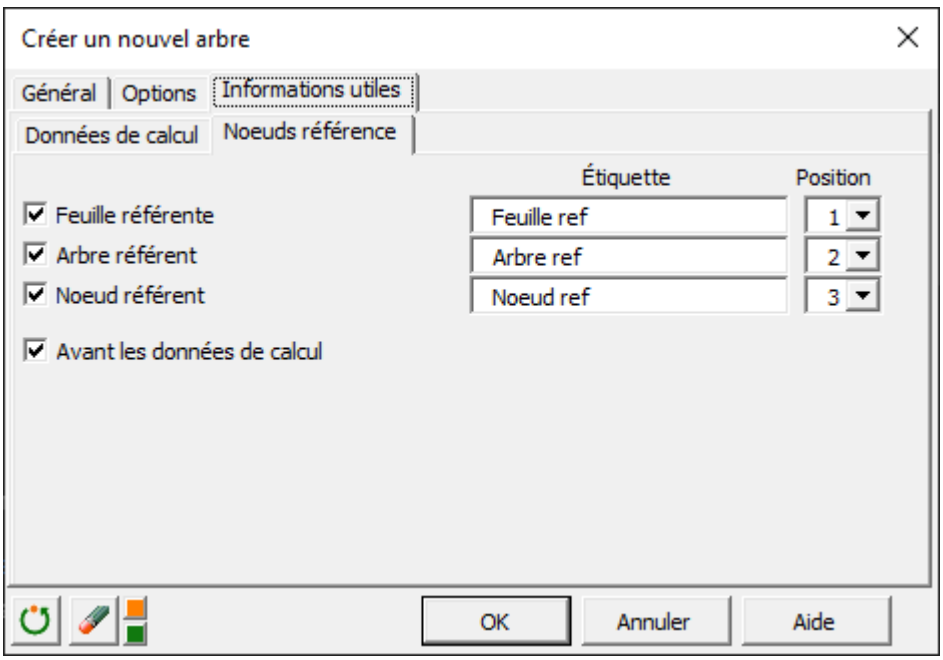
Onglet **Informations utiles - Données de calcul** :



- Vous choisissez ici les informations à afficher pour chaque bloc nœud. Leur définition se trouve dans la partie Description. Pour une information donnée, il est possible de modifier l'étiquette associée et sa position par rapport aux autres informations à afficher. Les informations relatives aux calculs avec utilité ne peuvent être sélectionnées que si l'option **R si fonction d'utilité** de l'onglet **Général** est également sélectionnée.

- **Nœuds décision, branche optimale pour** : Pour les nœuds décision, au niveau des blocs branche, il est possible d'afficher ou non si la branche est optimale pour le gain attendu et / ou pour le gain avec utilité attendu.

Onglet **Informations utiles - Nœuds référence** :

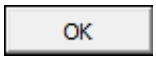


- Vous choisissez ensuite les informations relatives au nœud référent à afficher. Leur définition se trouve dans la partie Description. Pour une information donnée, il est

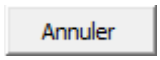
possible de modifier l'étiquette associée et sa position par rapport aux autres informations à afficher.

- **Avant les données de calcul** : cette option permet de choisir, pour chaque nœud référence, si les informations relatives au nœud référent sont affichées avant ou après les données de calcul.

Et pour finir :



: cliquez sur ce bouton pour valider vos paramètres et commencer à construire un nouvel arbre ou mettre à jour les paramètres d'un arbre existant.



: cliquez sur ce bouton pour fermer la boîte de dialogue sans créer de nouvel arbre ou modifier les paramètres d'un arbre existant.

: cliquez sur ce bouton pour afficher l'aide relative aux arbres de décision dans XLSTAT.


: cliquez sur ce bouton pour rétablir les options par défaut.

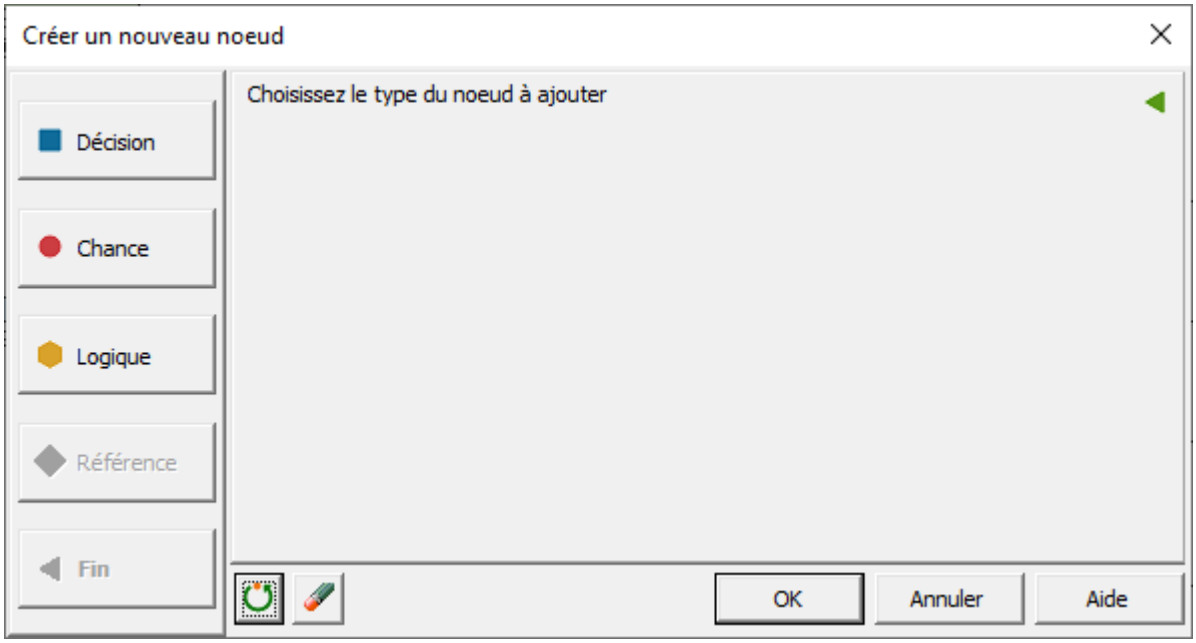

: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur le bouton orange pour enregistrer les paramètres de la boîte de dialogue dans un fichier ou sur le bouton vert pour charger les paramètres de la boîte de dialogue depuis un fichier.

Construction d'un nœud

Pour ajouter un nouveau nœud ou modifier un nœud existant, il suffit de cliquer sur l'icône d'un nœud. Cela déclench l'ouverture la boîte de dialogue de paramétrage d'un nœud. Cette boîte de dialogue peut être affichée et modifiée à tout moment. Le nœud cible se mettra à jour automatiquement si vous faites des modifications.

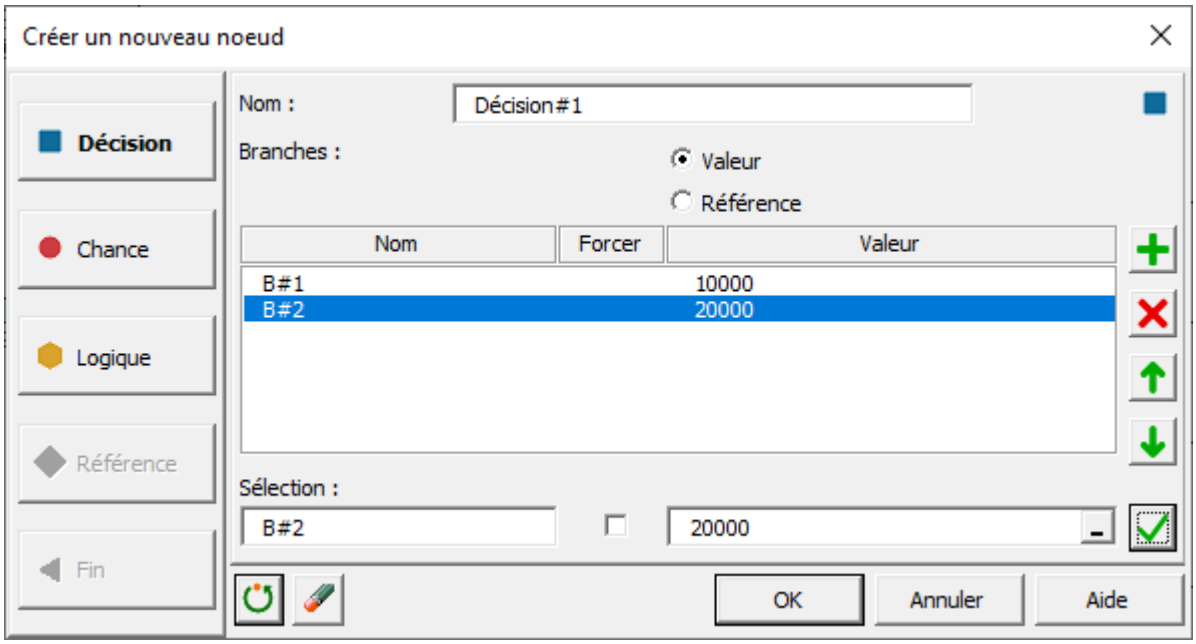
Lorsque vous cliquez sur l'icône d'un nœud final, une boîte de dialogue est affichée afin que vous sélectionniez le type de votre nouveau nœud.



Une fois le type choisi, la boîte de dialogue s'étend et vous permet de paramétrer votre nœud. Cette même boîte de dialogue étendue s'affiche lorsque vous cliquez sur l'icône d'un nœud existant, autre qu'un nœud final.

Nous allons maintenant détailler chacun des paramètres présents, selon le type de nœud.

Nœuds décision, chance et logique :



La partie gauche, avec les différents types de nœuds met à jour la partie droite selon le type du nœud choisi. Les nœuds décision, chance et logique ont un paramétrage équivalent car ce sont tous les trois des nœuds avec des branches. Nous allons donc les présenter ensemble.

- **Nom** : renseignez le nom du nœud.
- **Branches** :

C'est ici que les branches du nœud sont définies. Il est nécessaire de renseigner différents paramètres pour chaque branche :

- **Nom** : le nom de la branche.
- **Forcer** : cette option permet de forcer une branche à faire partie du chemin optimal. Une seule branche peut être forcée dans tout l'arbre.
- **Valeur** : le coût ou le profit associé à la branche. Si c'est un coût alors la valeur doit être négative. Ce paramètre peut faire référence à la valeur d'une cellule d'une feuille. Dans ce cas, sélectionnez la cellule concernée.
- **Probabilité** : la probabilité de réalisation associée à la branche. Ce paramètre n'est disponible que pour un nœud chance. Ce paramètre peut faire référence à la valeur d'une cellule d'une feuille. Dans ce cas, sélectionnez la cellule concernée.
- **Logique** : la formule logique associée à la branche. Ce paramètre n'est disponible que pour un nœud logique. Ce paramètre peut faire référence à la valeur d'une cellule d'une feuille. Dans ce cas, sélectionnez la cellule concernée. Il est aussi possible de directement écrire une formule ayant un résultat logique (VRAI ou FAUX).

Les options **Valeur** et **Référence** permettent de choisir le mode d'affichage des données pouvant faire référence à une cellule dans une feuille (Valeur, Probabilité et Logique). La liste des branches se met automatiquement à jour selon votre choix. Ces options ne s'appliquent qu'à la boîte de dialogue et non aux informations affichées dans un bloc branche.

Afin de modifier une branche, il suffit de la sélectionner dans la liste des branches, modifier ses paramètres dans la partie **Sélection** et mettre à jour dans la liste des branches en cliquant sur le bouton ✓.

D'autres actions sont possibles au niveau de la liste des branches :

- ✚ : cliquez sur ce bouton pour ajouter une nouvelle branche dans la liste des branches. Celle-ci sera ajoutée au-dessus de la branche sélectionnée.
- ✖ : cliquez sur ce bouton pour supprimer la branche sélectionnée de la liste des branches.
- ↑ : cliquez sur ce bouton pour déplacer vers le haut la branche sélectionnée dans la liste des branches.
- ↓ : cliquez sur ce bouton pour déplacer vers le bas la branche sélectionnée dans la liste des branches.

Et pour finir :

OK : cliquez sur ce bouton pour valider vos paramètres et commencer à construire un nouvel arbre ou mettre à jour les paramètres d'un arbre existant.

Annuler : cliquez sur ce bouton pour fermer la boîte de dialogue sans créer de nouvel arbre ou modifier les paramètres d'un arbre existant.

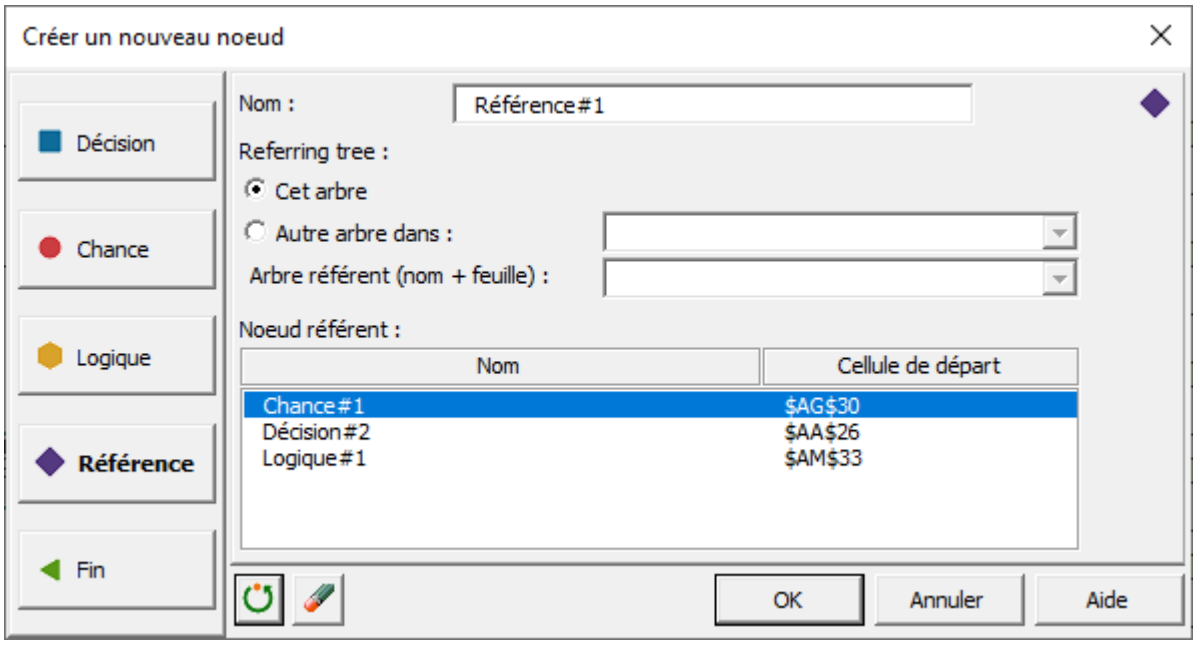
Aide : cliquez sur ce bouton pour afficher l'aide relative aux arbres de décision dans XLSTAT.

: cliquez sur ce bouton pour rétablir les options par défaut.

: cliquez sur ce bouton pour effacer les sélections de données.

: cliquez sur le bouton orange pour enregistrer les paramètres de la boîte de dialogue dans un fichier ou sur le bouton vert pour charger les paramètres de la boîte de dialogue depuis un fichier.

Nœud référence :



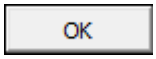
.

- **Nom** : renseignez le nom de votre nœud.
- **Arbre référent** :
  - **Cet arbre** : sélectionnez cette option si l'arbre référent est le même que l'arbre actif.
  - **Autre arbre dans** : sélectionnez cette option si l'arbre référent n'est pas l'arbre actif. Sélectionnez ensuite, dans la liste adjacente, le classeur dans lequel se trouve l'arbre référent.
  - **Arbre référent (nom + feuille)** : sélectionnez dans cette liste l'arbre référent. La liste est composée de tous les arbres présents dans le classeur précédemment sélectionné. Seul le nom de l'arbre s'affiche mais quand la liste est déroulée, le nom de la feuille où se trouve l'arbre est aussi renseignée pour plus de précisions.

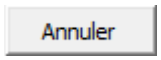
- **Nœud référent** :
    - **Nom** : nom de tous les nœuds présents dans l'arbre sélectionné dans la section précédente.
    - **Cellule de départ** : cellule de départ du bloc nœud.

  Sélectionnez dans la liste le nœud référent. Si celui-ci est dans le classeur actif, alors le bloc correspondant est sélectionné dans la feuille où il se trouve. Cela permet, si besoin, de mieux s'y retrouver.

Et pour finir :



: cliquez sur ce bouton pour valider vos paramètres et commencer à construire un nouvel arbre ou mettre à jour les paramètres d'un arbre existant.



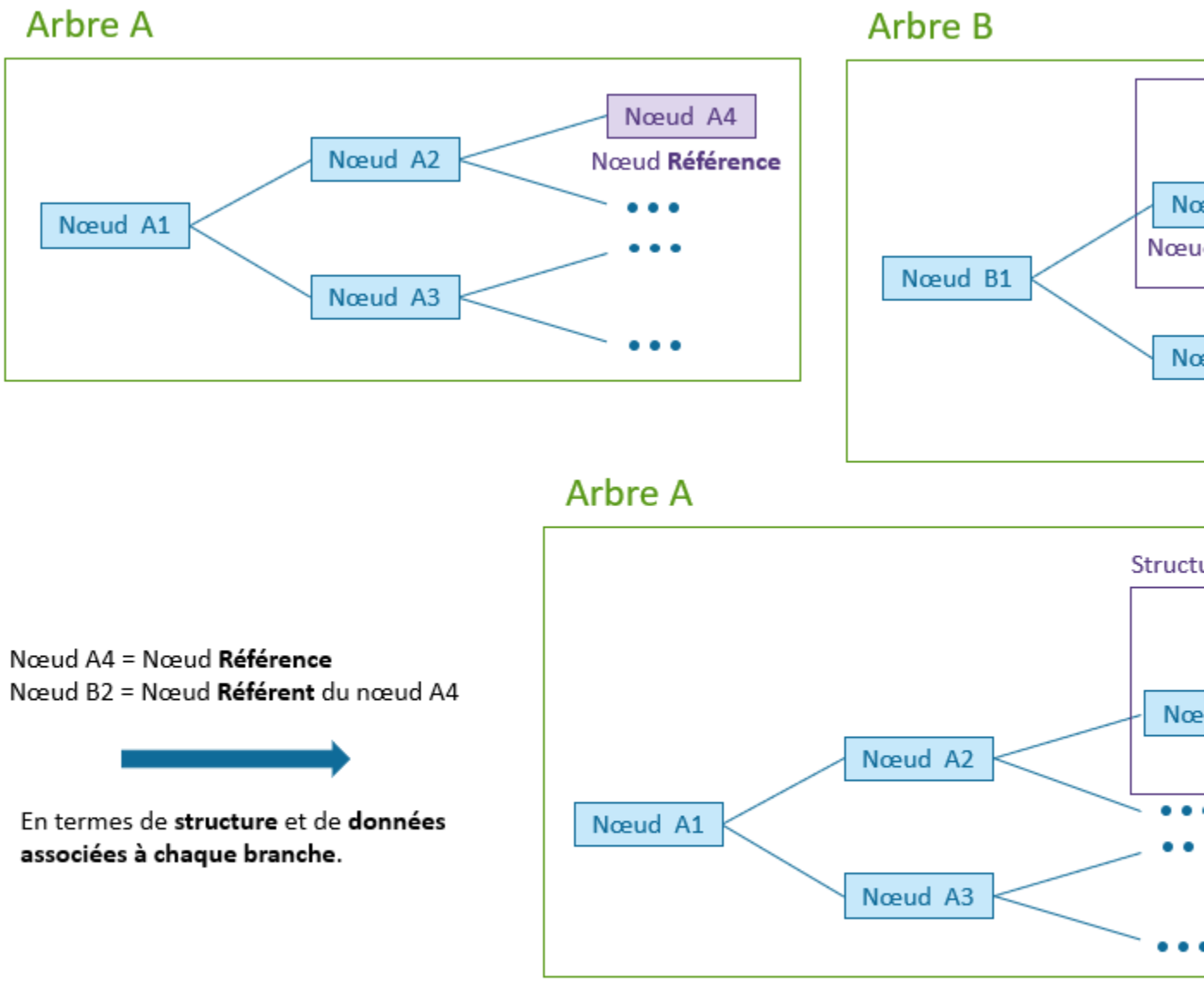
: cliquez sur ce bouton pour fermer la boîte de dialogue sans créer de nouvel arbre ou modifier les paramètres d'un arbre existant.

: cliquez sur ce bouton pour afficher l'aide relative aux arbres de décision dans XLSTAT.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.

Voici un schéma explicatif afin de mieux comprendre les notions de nœud référence et nœud référent :



Remarque : Si un arbre A contient un nœud référence dont le nœud référent est dans un arbre B alors l'arbre B ne pourra pas avoir de nœud référent dans l'arbre A. Il faut avoir les deux nœuds référence dans l'arbre A.

Actions sur un arbre

Directement dans les blocs constituant l'arbre

Il est possible de modifier certaines informations relatives à un arbre, directement dans la feuille Excel où il se trouve. Les cellules concernées sont colorées. Voici la liste de ces informations :

- Nom de l'arbre

- Nom d'un nœud
- Nom d'une branche
- Valeur d'une branche
- Probabilité d'une branche (dans le cas d'un nœud chance)
- Formule logique d'une branche (dans le cas d'un nœud logique)

Clic sur l'icône d'un nœud

- Décision :
- Chance :
- Logique :
- Référence :
- Final :

Si vous cliquez sur l'icône d'un nœud alors la boîte de dialogue permettant de le paramétrer va s'ouvrir. Vous pouvez ainsi modifier son paramétrage (même son type) et le valider en cliquant sur le bouton Continuer.

Clic droit sur le bloc arbre

Si vous faites un clic droit sur le bloc arbre alors vous pourrez avoir accès à un menu XLDTREE :

Voici la liste des différentes actions possibles :

- **Ouvrir la boîte de dialogue de paramétrage de l'arbre sélectionné** : cliquez sur cet élément si vous souhaitez afficher la boîte de dialogue de paramétrage de l'arbre.
- **Mettre en surbrillance le chemin optimal pour l'arbre sélectionné** : cliquez sur cet élément si vous souhaitez mettre en valeur le chemin optimal pour l'arbre entier. La notion de chemin optimal est abordée dans la partie Calculs et chemin optimal.
- **Enlever le chemin optimal pour l'arbre sélectionné** : cliquez sur cet élément si vous ne souhaitez plus mettre en valeur le chemin optimal pour l'arbre entier.
- **Supprimer l'arbre sélectionné** : cliquez sur cet élément si vous souhaitez supprimer l'arbre proprement.

Clic droit sur un bloc nœud

Si vous faites un clic droit sur un bloc nœud alors vous pourrez avoir accès à un menu XLDTREE : 

Voici la liste des différentes actions possibles :

- **Créer un nouveau nœud** : cliquez sur cet élément, disponible depuis un nœud final, si vous souhaitez afficher la boîte de dialogue de paramétrage d'un nœud et ainsi créer un nouveau nœud à la place du nœud final sélectionné.
- **Ouvrir la boîte de dialogue de paramétrage du nœud sélectionné** : cliquez sur cet élément si vous souhaitez afficher la boîte de dialogue de paramétrage d'un nœud afin de mettre à jour le nœud sélectionné.
- **Mettre en surbrillance le chemin optimal à partir du nœud sélectionné** : cliquez sur cet élément si vous souhaitez mettre en valeur le chemin optimal à partir de ce nœud. La notion de chemin optimal est abordée dans la partie Calculs et chemin optimal.
- **Enlever le chemin optimal commençant au nœud sélectionné** : cliquez sur cet élément si vous ne souhaitez plus mettre en valeur le chemin optimal à partir de ce nœud.
- **Insérer un nouveau nœud avant le nœud sélectionné** : cliquez sur cet élément si vous souhaitez insérer un nouveau nœud avant le nœud cible. Par défaut, c'est un nœud décision avec deux branches qui va être inséré. Le nœud cible sera sur la première branche de ce nouveau nœud.
- **Supprimer le sous-arbre à partir du nœud sélectionné** : cliquez sur cet élément si vous souhaitez supprimer le nœud cible ainsi que tous ses nœuds enfants. Le nœud cible sera remplacé par un nœud final.
- **Copier le sous-arbre à partir du nœud sélectionné** : cliquez sur cet élément si vous souhaitez copier le sous-arbre composé du nœud cible avec tous ses nœuds enfants.
- **Coller le sous-arbre à la place du nœud sélectionné et de ses enfants** : cliquez sur cet élément si vous souhaitez remplacer le sous-arbre composé du nœud cible et de ses nœuds enfants par le sous-arbre composé du nœud copié précédemment avec ses nœuds enfants.

Clic droit sur un bloc branche

Si vous faites un clic droit sur un bloc branche alors vous pourrez avoir accès à un menu XLDTREE : 

Voici la liste des différentes actions possibles :

- **Ajouter une nouvelle branche au-dessus de la branche sélectionnée** : cliquez sur cet élément si vous souhaitez ajouter une nouvelle branche. Elle sera ajoutée au-dessus de la branche cible.

- **Supprimer la branche sélectionnée et ses enfants** : cliquez sur cet élément si vous souhaitez supprimer la branche cible avec tous ses enfants.
- **Déplacer la branche sélectionnée vers le haut** : cliquez sur cet élément si vous souhaitez déplacer la branche cible vers le haut (dans le cas d'un arbre affiché horizontalement).
- **Déplacer la branche sélectionnée vers le bas** : cliquez sur cet élément si vous souhaitez déplacer la branche cible vers le bas (dans le cas d'un arbre affiché horizontalement).
- **Déplacer la branche sélectionnée vers la gauche** : cliquez sur cet élément si vous souhaitez déplacer la branche cible vers la gauche (dans le cas d'un arbre affiché verticalement).
- **Déplacer la branche sélectionnée vers la droite**: cliquez sur cet élément si vous souhaitez déplacer la branche cible vers la droite (dans le cas d'un arbre affiché verticalement).
- **Forcer la branche sélectionnée à être sur le chemin optimal** : cliquez sur cet élément si vous souhaitez forcer le chemin optimal à passer par la branche cible. Cela peut être utile lorsque vous revenez plus tard sur un arbre existant, à un moment où certains choix ont déjà été faits ou réalisés. Une seule branche, parmi toutes celles de l'arbre, peut-être forcée. Si une autre branche était déjà forcée alors elle ne le sera plus. La branche forcée a son nom affiché dans une couleur différente de celle des autres noms de branches. La notion de chemin optimal est abordée dans la partie Calculs et chemin optimal.
- **Arrêter de forcer la branche sélectionnée à être sur le chemin optimal** : cliquez sur cet élément si vous souhaitez arrêter de forcer le chemin optimal à passer par la branche cible. Plus aucune branche parmi toutes celles de l'arbre ne sera donc forcée.

Calculs et chemin optimal

Calculs

Pour chaque bloc d'un arbre, que ce soit le bloc de l'arbre lui-même, un bloc nœud ou un bloc branche, des calculs sont effectués à chaque modification. Leur résultat est affiché directement sous forme de formules, dans les blocs concernés. Contrairement aux informations modifiables dont la cellule est colorée, les informations résultats du calcul d'une formule sont sur fond blanc. Il est possible de remettre en place des formules supprimées par erreur en générant l'arbre à nouveau. Il suffit d'ouvrir la boîte de dialogue de paramétrage de l'arbre et de cliquer sur le bouton OK.

Comme vu précédemment il est possible de choisir quelles sont les informations affichées pour les blocs nœuds à partir de la boîte de dialogue de paramétrage d'un arbre.

Chemin optimal

Cet outil n'est pas disponible sous Mac pour l'instant.

Le chemin optimal représente le chemin répondant le mieux au mode de calcul choisi. Il dépend du gain attendu, avec utilité ou non, de chaque nœud. Le choix d'un chemin optimal pour le gain attendu ou le gain avec utilité attendu, se fait via la boîte de dialogue de paramétrage d'un arbre. Le comportement est différent selon le type de nœud :

- **Nœud décision** : le chemin optimal passe par la branche dont le gain, avec utilité ou non, répond le mieux au mode de calcul choisi.
- **Nœud chance** : le chemin optimal arrivant sur un nœud chance, passe par toutes les branches de ce nœud. En effet, il n'est pas possible de savoir à l'avance quelle branche sera réalisée, peu importe sa probabilité de réalisation. Le gain, avec utilité ou non, tient compte de ce comportement puisqu'il est la somme pondérée (par la probabilité de réalisation) du gain, avec utilité ou non, attendu du nœud enfant de chaque branche.
- **Nœud logique** : le chemin optimal passe par la seule branche ayant un résultat VRAI. Comme déjà vu, il n'est pas possible d'avoir plusieurs branches avec un résultat VRAI. Si tel est le cas alors le gain, avec utilité ou non, et donc le chemin optimal, ne peuvent être calculés et sont en erreur. Si toutes les branches ont un résultat FAUX alors le gain, avec utilité ou non, et donc le chemin optimal, ne sont pas disponibles pour ce nœud.
- **Nœud référence** : un nœud référence a pour référent un nœud décision, chance ou logique. Il suffit donc de se reporter à ces types de nœuds pour connaître le comportement du chemin optimal au niveau de ce type de nœud. Cependant, ce comportement restera invisible car le nœud référence d'un arbre n'affiche pas ses nœuds enfants.

Le chemin optimal, une fois activé, est mis en évidence sur l'arbre. Il peut concerner l'arbre entier ou bien ne commencer qu'à partir d'un nœud que vous aurez choisi par clic droit. Tant que l'arbre optimal est activé, il est automatiquement recalculé et son affichage mis à jour, à chaque modification de l'arbre.

Exemples

Des exemples d'utilisation de l'outil Arbres de décision sont disponibles sur le Centre d'aide XLSTAT :

http://www.xlstat.com/demo-dct_fr.htm

http://www.xlstat.com/demo-dct2_fr.htm

http://www.xlstat.com/demo-dct3_fr.htm

Réseaux Bayésiens

Le module Réseaux Bayésiens de XLSTAT permet l'analyse statistique au moyen d'un réseau bayésien. Très populaire en intelligence artificielle, les réseaux bayésiens permettent notamment de représenter une connaissance et ses incertitudes. C'est un outil d'aide à la décision dont la fonction principale est de faire apparaître des relations de causalité entre des variables.

Les réseaux bayésiens sont utilisés dans les domaines de la finance, par exemple pour analyser des risques de fraudes à la carte bancaire, du médical pour l'aide au diagnostic, ou de l'industrie.

Dans cette section :

Description

Un réseau bayésien est un outil statistique permettant de modéliser les relations de dépendance ou d'indépendance conditionnelle entre des variables aléatoires. Cette méthode est apparue suite aux recherches pionnières de Judea Pearl en 1988 sur le développement des techniques utilisées en intelligence artificielle.

L'originalité de cette méthode est qu'elle offre un cadre formel pour représenter la structure relationnelle des variables du réseau. D'un côté nous disposons d'une description qualitative avec un graphique et de l'autre côté une description quantitative avec des lois de probabilités. Le graphique est un moyen de schématiser le réseau avec des noeuds et des arcs les reliant, ce qui facilite la compréhension du problème et l'interprétation des résultats.

Dans un réseau bayésien, le graphe doit être orienté et respecter la règle dite "acyclique" c'est à dire qu'il ne possède pas de circuit (voir figure 1). On parle alors de graphe acyclique dirigé (DAG pour *directed acyclic graph*). Lorsqu'un arc se dirige d'un noeud A vers un noeud B, on dira que le noeud A est un parent du noeud B et que le noeud B est un noeud enfant du noeud A. Les lois de probabilités sont données pour chaque noeud suivant le statut de ce dernier dans le réseau (enfant ou parent).

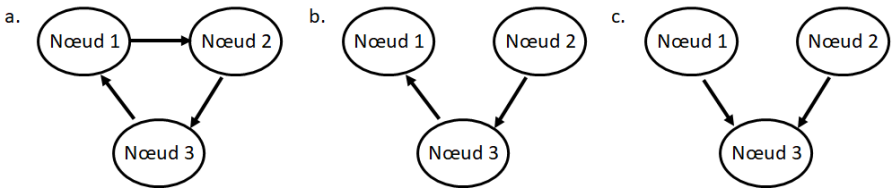


Figure 1. Exemple de graphes orientés à 3 noeuds. Le graphe a. est un circuit alors que les graphes b. et c. sont acycliques.

Le réseau bayésien ainsi défini sert ensuite à répondre à des requêtes, c'est à dire, à évaluer le modèle et obtenir des probabilités a posteriori étant donné de nouvelles informations. soient U l'ensemble des noeuds constituant le réseau et $P(U)$ la distribution de probabilités sur cet ensemble. Si nous disposons d'une nouvelle information ϵ sur une ou plusieurs variables, alors on souhaite mettre à jour la connaissance que représente le réseau bayésien à travers $P(U)$ pour cette nouvelle information. Cette mise à jour est appelée l'inférence. On peut en dissocier 2 types :

- d'effet à cause : la nouvelle information vient d'un noeud enfant et se propage vers son ou ses noeuds parents. Dans cette situation le réseau bayésien sert à établir un diagnostic.
- de cause à effet : la nouvelle information vient d'un noeud parent et se propage vers ses noeuds enfants. Dans cette situation le réseau bayésien sert à faire des simulations ou des prédictions.

Mathématiquement parlant, l'inférence dans un réseau bayésien est le calcul de $P(U|\epsilon)$, c'est-à-dire le calcul de la probabilité a posteriori du réseau sachant ϵ. Deux équations sont fondamentales dans ce calcul : le théorème de Bayes qui s'écrit selon l'équation suivante :

$$P(A|B) = \frac{P(A|B)P(A)}{P(B)},$$

où A et B sont deux variables aléatoires et la formule de la probabilité jointe ci-dessous :

$$P(A,B) = P(A|B)P(B) = P(B|A)P(A).$$

Le terme $P(A|B)$ est appelé *probabilité conditionnelle de A sachant B* (ou encore de A sous condition B). Le terme $P(B|A)$, pour un A connu, est appelé la fonction de vraisemblance de B. Les termes $P(A)$ et $P(B)$ sont respectivement la probabilité a priori de A ou la probabilité marginale de A et la probabilité marginale ou a priori de B.

Cependant, ce calcul peut se révéler complexe. En effet, selon la complexité du réseau (plus la topologie du réseau est simple, plus l'inférence est aisée), le calcul, par exemple de la probabilité jointe, qui revient à faire le produit des probabilités conditionnelles et marginales pour toutes les valeurs des variables du réseau devient coûteux en temps de calcul. Il a même été démontré par G. F. Cooper (1990) que, dans le cas général, il s'agit d'un problème NP-difficile.

C'est de là que sont nés plusieurs algorithmes permettant le calcul dans un système complexe probabilisé. XLSTAT utilise l'algorithme d'inférence exacte connu sous le nom d'arbre de jonction (Jensen et al., 1990). Son principe est de transformer le graphe en un arbre puis de lui associer son arbre de jonction. Ce dernier est aussi un graphe dont chaque noeud forme une clique. En théorie des graphes, une clique est un ensemble de sommets deux-à-deux adjacents c'est à dire formée d'un sous-ensemble de noeuds du graphe original complètement connectés. Dans l'exemple donné sur la figure 2, le graphe b est formé de 4 cliques toutes composées de 2 noeuds du graphe original a. La probabilité jointe du réseau devient ainsi le produit des probabilités jointes de chaque clique.

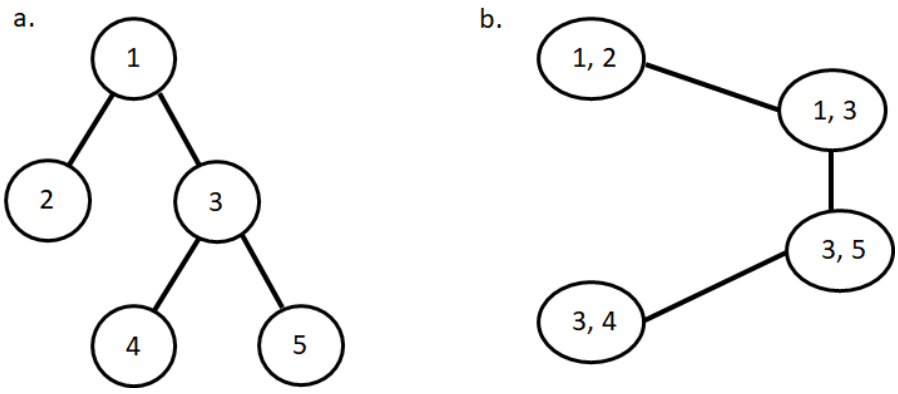


Figure 2. Exemple d'un arbre (a.) et d'un arbre de jonction (b.).

La démarche à suivre pour analyser un réseau bayésien dans XLSTAT est la suivante :

1. Ouverture d'un projet : dans le menu XLSTAT allez dans le menu Réseaux bayésiens et ouvrez un nouveau projet. Une fenêtre s'ouvre vous proposant le choix entre le mode classique et le mode expert (voir le paragraphe Options et sélection d'objet sur le graphe pour plus d'information. Dans les deux cas un classeur s'ouvrira composé de plusieurs feuilles qui seront utilisées pour créer votre réseau (voir la description dans la section projets).

2. Construction d'un graphe : dessinez votre réseau à l'aide des boutons fournis et nommez chacun de vos noeuds. Pour plus d'information sur cette étape allez directement à la

section Construction d'un graphe.

3. Définition des distributions de probabilités : pour chaque variable vous devez remplir un tableau de probabilités. Ceci peut se faire de deux manières : en mode expert, c'est à dire à partir des connaissances humaines, ou en mode apprentissage pour lequel le réseau bayésien peut être appris automatiquement à partir de données. Ces deux options sont décrites dans la section Définition des tableaux de probabilités.

4. Inférence : vous pouvez à présent faire des requêtes sur votre réseau bayésien et lancer les calculs de probabilités à partir du bouton dédié. Pour plus d'information sur l'utilisation de cette fonctionnalité, lisez la section Analyse d'un réseau bayésien.

Projets

Les projets Réseaux Bayésiens sont des classeurs Excel particuliers. Lorsque vous créez un nouveau projet, son nom par défaut commence par BNBook. Vous pouvez ensuite le sauvegarder sous un nom de votre choix. Le bouton "Enregistrer" dans les options proposées du menu Réseaux Bayésiens vous permet de sauvegarder votre projet en utilisant l'extension *.xlsm.

Un projet brut Réseaux Bayésiens contient toujours au départ deux feuilles puis trois lorsque les tableaux de probabilités sont définis. Ces feuilles, qui ne doivent pas être supprimées, sont :

- Data : c'est une feuille Excel vide, dans laquelle vos données doivent être copiées/collées.
- BNGraph : c'est une feuille contenant une zone de dessin, vide au départ, et une barre d'outils dont une description est faite dans la section suivante. Elle doit être utilisée pour dessiner le graphe du réseau bayésien.
- Tableaux de probabilités : c'est une feuille qui contient les distributions de probabilités de tous les noeuds qui ont été dessinés sur le graphe de la feuille BNGraph. Elles sont formalisées sous la forme d'un tableau par noeud. Pour chaque tableau le nombre de colonnes est celui du nombre de parents du noeud plus une colonne pour le noeud lui-même et plus une colonne pour les probabilités. Un tableau a au minimum deux colonnes. On a deux colonnes pour un noeud marginal (il n'a pas de noeud parent). Le nombre de lignes est défini par le nombre de combinaisons des modalités du noeud et de ses noeuds parents. Par exemple, soient 3 noeuds A, B et C avec respectivement 2, 3 et 2 modalités dont les intitulés sont $\{a_1, a_2\}$, $\{b_1, b_2, b_3\}$, $\{c_1, c_2\}$. Les noeuds B et C sont les parents du noeud A. Le tableau de probabilités du noeud A se présente sous la forme suivante :

A	B	C	Probabilités
a_1	b_1	c_1	----------
a_2	b_1	c_1	----------
a_1	b_2	c_1	----------
a_2	b_2	c_1	----------
a_1	b_3	c_1	----------
a_2	b_3	c_1	----------
a_1	b_1	c_2	----------
a_2	b_1	c_2	----------
a_1	b_2	c_2	----------
a_2	b_2	c_2	----------
a_1	b_3	c_2	----------
a_2	b_3	c_2	----------

L'ordre des parents est décroissant avec le nombre de modalités. L'ordre des modalités dans le tableau est l'ordre donné par l'utilisateur.

Une fois que le réseau bayésien est dessiné et les tableaux de probabilités remplis, vous pouvez lancer les calculs de probabilités. Les résultats sont affichés dans une feuille Excel, à la suite de la feuille BNGraph.

Il est possible d'enregistrer un modèle avant de le modifier, afin de pouvoir éventuellement le modifier par la suite (voir la section Options et sélection d'objet sur le graphe pour plus de détails).

Barre d'outils

Une barre d'outils est disponible en haut à gauche de la feuille BNGraph qui comprend huit boutons orange alignés horizontalement :

Ces boutons sont utiles dans la construction du graphe, dans la définition des tableaux de probabilités et l'analyse du réseau bayésien. Plus précisement les quatre premiers boutons sont dédiés au dessin du réseau bayésien, les deux suivants à renseigner la table de probabilités, le septième sert au calcul des probabilités du réseau bayésien dessiné et le dernier bouton redirige vers cette aide. Le mode d'utilisation et la fonctionnalité de l'ensemble de ces boutons sont décrits dans les sections suivantes de cette aide.

La barre d'outils est uniquement visible lorsque vous êtes sur la feuille BNGraph.

Options et sélection d'objet sur le graphe

Le module Réseaux Bayésiens de XLSTAT vous propose 3 options :

 Cliquez sur ce bouton pour ouvrir un nouveau projet Réseaux Bayésiens. Une fenêtre s'ouvre alors vous proposant le choix entre un affichage de la méthode en mode classique ou en mode expert. Le mode classique est privilégié lorsque vous disposez d'un jeu de données et que vous souhaitez que les distributions de probabilités soient automatiquement calculées. Le mode expert est lui destiné comme son nom l'indique aux "experts" de la problématique étudiée car il permet à l'utilisateur de définir lui-même les modalités et les probabilités de chaque variable. Dans les deux cas un nouveau projet s'ouvre dont vous trouverez une description dans la section précédente.

 Cliquez sur ce bouton pour ouvrir un projet Réseau Bayésien existant.

 Cliquez sur ce bouton pour enregistrer le projet Réseau Bayésien actif. Ce bouton n'est accessible que si des modifications ont été effectuées dans le projet.

La sélection des objets (noeud ou arc) sur la feuille BNGraph se fait uniquement au moyen de la touche Ctrl + clic gauche avec la souris, ou cmd + click gauche pour les utilisateurs Mac. Il ne faut surtout pas utiliser le clic droit de la souris et le menu déroulant d'Excel.

La déselection se fait avec la touche Echap et s'applique à tous les objects sélectionnés.

Vous pouvez bouger les objets avec la souris en s'assurant au préalable d'avoir tout déselectionné puis en sélectionnant les objets souhaités. Il est possible d'utiliser également les flèches du clavier.

Vous pouvez supprimer un noeud ou un arc, un seul à la fois seulement, en le sélectionnant puis en appuyant sur la touche Suppr de votre clavier. Lorsque vous reliez deux noeuds par un arc, ce dernier ne doit pas être bougé pour relier deux autres noeuds car il a été défini pour les deux premiers noeuds uniquement.

Du texte grisé à gauche de la feuille de dessin BNGraph, sous la ligne de boutons, résume les raccourcis clavier disponibles qui sont :

- Sélection de noeud ou d'arc : Ctrl + click gauche (ou cmd + click gauche pour les utilisateurs Mac)
- Désélectionner tout : Echap
- Bouger le noeud : Désélectionner tout + sélection de noeud + bouger avec la souris
- Supprimer : Ctrl + Suppr
- Relier 2 noeuds : Ctrl + A
- Changer la direction de l'arc : Ctrl + D
- Editeur de modalités : Ctrl + E (visible uniquement en mode expert)
- Editeur de données : Ctrl + L (visible uniquement en mode classique)

Construction d'un graphe

Le graphe d'un réseau bayésien est matérialisé par des noeuds et des arcs qui relient les noeuds. Dans un projet Réseaux bayésiens (voir projets) le graphe doit être dessiné sur la feuille nommée BNGraph à partir de la barre d'outils. La sélection des noeuds et des arcs se fait selon la description faite dans la section Options et sélection d'objet sur le graphe.

Cliquez sur ce bouton pour ajouter un noeud. Une fois le bouton actionné, celui-ci devient gris indiquant qu'il est actif, vous pouvez dès lors positionner le noeud sur la feuille de dessin, là où vous le souhaitez. Au positionnement une fenêtre s'ouvre pour nommer le noeud. Vous avez la possibilité de modifier le nom en cliquant à nouveau sur le noeud ou de le renommer plus tard .

Cliquez sur ce bouton pour nommer un noeud. Pour cela, sélectionnez d'abord un noeud puis cliquez sur ce bouton. Vous pouvez nommer qu'un seul noeud à la fois.

Cliquez sur ce bouton pour lier deux noeuds avec un arc. Pour l'utiliser vous devez d'abord sélectionner le noeud parent puis le noeud enfant et enfin le bouton Arc. Vous ne pouvez pas relier plus de deux noeuds à la fois et vous pouvez créer qu'un seul arc entre deux même noeuds.

Cliquez sur ce bouton pour changer la direction d'un arc entre deux noeuds. Pour l'utiliser, vous devez d'abord sélectionner l'arc de votre choix, puis cliquez sur le bouton. Vous ne pouvez pas changer la direction de plusieurs arcs à la fois.

Définition des tableaux de probabilités

Les variables impliquées dans le réseau bayésien peuvent être qualitatives ou quantitatives. Elles doivent comprendre au minimum 2 modalités. Les tableaux de probabilités indiquentpour ces modalités les différentes valeurs prises. Lorsqu'une variable est dépendante d'une ou de plusieurs variables le tableau des probabilités donne les valeurs prises pour toutes les combinaisons de modalités de l'ensemble des variables.

Dans l'application Réseaux bayésiens on peut définir les tableaux de probabilités de deux façons :

- en mode classique,
- en mode expert.

L'utilisation de ces deux modes est expliquée dans les deux paragraphes suivants. Dans les deux cas les tableaux de probabilités sont affichés dans une feuille dédiée, comme décrit dans la section projets. Le bouton  , plus utile en mode expert, est commun aux deux modes. Il permet de visualiser et/ou de modifier la valeur d'une probabilité spécifique. Vous trouverez une description détaillée sur son utilisation dans le paragraphe expert.

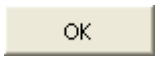
En mode classique

Les modalités et les tableaux de probabilités sont automatiquement définis à partir d'un jeu de données. Le libellé des colonnes doit contenir le nom des variables. Attention à bien veiller à ce que leur nom soit le même que celui utilisé pour nommer les noeuds de votre réseau.

L'algorithme fonctionne avec des variables qualitatives. Lorsque des variables quantitatives sont sélectionnées celles-ci sont transformées en variables qualitatives de la façon suivante. D'abord, pour chacune d'elle, leurs valeurs sont triées en ordre croissant, puis discrétisées en un nombre maximal de dix intervalles. Ces nouveaux intervalles sont ensuite utilisés pour recoder les valeurs et deviennent les modalités. Un tableau des modalités est ensuite calculé pour toutes les variables en tenant compte de leur structure relationnelle. Ce tableau ressemble à celui présenté dans le paragraphe projets sans la colonne probabilité. La fréquence d'apparition de ces modalités dans les données est ensuite calculée, pour finalement être convertie en probabilité avec une somme marginale (somme des valeurs prises par une variable pour une valeur donnée des autres variables) égale à 1. Lorsque la fréquence est nulle pour plus de la moitié des modalités d'une variable, les probabilités ne sont pas calculées par l'algorithme mais remplacées par une valeur manquante dans la feuille de résultats. A l'inverse, lorsque la fréquence est nulle pour moins de la moitié alors les probabilités sont calculées comme la moyenne de la différence de 1 et les probabilités déjà calculées.

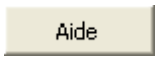
Pour utiliser cette fonctionnalité, cliquez sur le bouton suivant pour charger l'ensemble de vos données :

Cette action entraîne l'affichage d'une boîte de dialogue composée de plusieurs onglets correspondant aux différentes options disponibles, tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.



: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.


: cliquez sur ce bouton pour changer la façon dont vous chargez les données dans XLSTAT. Si le bouton possède une icône de souris, XLSTAT vous permet de faire une sélection à la souris des données. Si le bouton affiche une icône de liste, XLSTAT affichera une liste des variables disponibles que vous pouvez sélectionner. Si le bouton possède une icône orange, des boutons complémentaires avec un point d'interrogation  sont affichés pour vous permettre d'importer les données à partir de fichiers.

Onglet **Général** :

Données qualitatives : activez cette option si vous souhaitez utiliser des variables qualitatives puis sélectionnez ces variables.

Données quantitatives : activez cette option si vous souhaitez utiliser des variables quantitatives puis sélectionnez ces variables.

Poids : activez cette option si vous voulez pondérer les observations. Si vous n'activez pas cette option, les poids seront tous considérés comme valant 1. Les poids doivent être impérativement supérieurs ou égaux à 0.

Onglet **Données manquantes** :

Ne pas accepter les valeurs manquantes : activez cette option pour que XLSTAT empêche la poursuite des calculs si des valeurs manquantes sont détectées.

Supprimer les observations : activez cette option pour supprimer les observations comportant des données manquantes.

Estimer les données manquantes : activez cette option pour estimer les données manquantes avant le début des calculs.

- **Moyenne ou mode** : activez cette option pour estimer les données manquantes en utilisant la moyenne (variables quantitatives) ou le mode (variables qualitatives) pour les

variables correspondantes.

- **Plus proche voisin** : activez cette option pour estimer les données manquantes d'une observation en recherchant le plus proche voisin de l'observation.

Onglet **Sorties** :

Les résultats sont affichés dans une nouvelle feuille nommée **tableaux de probabilités**. Une synthèse des données est d'abord présentée au début de la feuille en précisant le nombre de données manquantes, la valeur de remplacement si cette option a été choisie et les statistiques descriptives des variables. Viennent ensuite les distributions de probabilités de chaque variable en fonction du statut de cette dernière dans le réseau bayésien. Elles sont présentées sous forme de tableau comme celui de la section projets. En fin de feuille, vous trouverez un bouton pour lancer directement les analyses sur le réseau bayésien dessiné à partir de ces tableaux de probabilités.

En mode expert

C'est l'utilisateur qui définit lui-même, pas à pas, toutes les informations requises sur les variables, à savoir les modalités et les probabilités. Pour cela deux boutons s'offrent à vous pour l'une et l'autre action.

 Cliquez sur ce bouton pour définir les modalités de vos variables. Une fenêtre s'ouvre vous permettant de sélectioner les données de la feuille Data (voir la section projets). Dans cette feuille sont listées en colonne les modalités des variables, avec une colonne par variable. L'en-tête des colonnes contient le nom des variables et doit correspondre au nom des noeuds dessinés sur le graphe BNGraph. Vous pouvez sélectionner une ou plusieurs colonnes. Si vous souhaitez définir les modalités d'une seule variable, vous pouvez présélectionner le noeud correspondant sur le graphe BNGraph, le nom du noeud sélectionné apparaîtra alors dans la fenêtre de l'éditeur de modalités. Lorsque vous cliquez sur Ok, les tableaux de probabilités des variables sélectionnées sont mis à jour dans la feuille dédiée.

 Cliquez sur ce bouton pour afficher et/ou modifier les probabilités d'une variable. Pour cela vous devez d'abord sélectionner un noeud sur le graphe BNGrpah (voir la section Options et sélection d'objet sur le graphe). Une fenêtre s'ouvre alors et affiche le tableau de probabilités de la variable sélectionnée. Vous pouvez modifier la valeur d'une probabilité à la fois. Pour cela, vous devez la sélectionner puis cliquer sur le bouton "éditer". Cette fois-ci une fenêtre s'ouvre dans laquelle vous pouvez saisir la nouvelle valeur ou la garder ainsi. En cliquant sur OK, cette valeur apparait dans le tableau des probabilités affiché. Si vous cliquez sur OK à nouveau cette valeur sera sauvegardée dans la feuille excel regroupant les tableaux de probabilités de toutes les variables.

Analyse d'un réseau bayésien

Une fois le modèle conçu sur la feuille BNGraph et une fois que toutes les probabilités ont été définies pour chaque variable, vous pouvez cliquer sur le bouton  de la barre d'outils pour lancer les analyses sur le réseau bayésien. Cela est aussi possible à partir du bouton situé à la fin des tableaux de probabilités. Une boîte de dialogue s'affiche avec les onglets suivants.

Onglet **Général** :

Source de données : sélectionnez une feuille contenant les tableaux de probabilités générés par XLSTAT (voir la section Définition des tableaux de probabilités) que vous souhaitez analyser.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Onglet **Données manquantes** :

Ne pas accepter les valeurs manquantes : activez cette option pour que XLSTAT empêche la poursuite des calculs si des valeurs manquantes sont détectées.

Estimer les données manquantes : activez cette option pour estimer les données manquantes avant le début des calculs.

- **Moyenne** : activez cette option pour estimer les données manquantes en utilisant la moyenne.

Onglet **Sorties** :

Probabilités marginales : activez cette option pour afficher la distribution des probabilités marginales de chaque noeud/variable.

Distribution de probabilité jointe : activez cette option pour afficher la distribution de probabilité jointe de chaque clique.

Probabilités conditionnelles : activez cette option pour afficher la distribution des probabilités conditionnelles de chaque noeud/variable.

Onglet **Graphiques** :

Graphique des probabilités marginales : activez cette option pour afficher le graphique de la distribution des probabilités marginale de chaque noeud/variable.

Graphique des probabilités conditionnelles : activez cette option pour afficher le graphique de la distribution de probabilités de chaque noeud/variable.

Résultats

Les résultats obtenus répondent à l'ensemble des requêtes possibles sur le réseau bayésien ainsi dessiné et renseigné des tableaux de probabilités.

Distribution de probabilité marginale de chaque noeud : ce résultat, sous forme de tableau, correspond aux probabilités marginales de chaque noeud dessiné sur le graphe BNGraph. Si l'option graphique est sélectionnée le résultat est aussi affiché sous forme d'un diagramme en bâton en dessous de chaque tableau.

Distribution de probabilité jointe pour chaque clique : Ce résultat présente le nombre de cliques calculé, la liste des noeuds impliqués dans chacune des cliques, puis le tableau de distribution de probabilité jointe pour chaque clique.

Distribution de probabilité conditionnelle de chaque noeud : ce résultat, sous forme de tableau, correspond aux probabilités conditionnelles de chaque noeud dépendant (Elles sont schématisées par un arc sur le graphe BNGraph). Si l'option graphique est sélectionnée le résultat est aussi affiché sous forme d'un diagramme en bâton en dessous de chaque tableau.

Exemple

Des tutoriels sur l'utilisation du module XLSTAT-Réseaux bayésiens sont disponibles sur le Centre d'aide XLSTAT :

http://www.xlstat.com/demo-bynf.htm

Bibliographie

Cooper, G. (1990). Computational complexity of probabilistic inference using bayesian belief networks. *Artificial Intelligence*, **42**, 393-405.

Jensen, F.V. (1996). An introduction to Bayesian Networks. Taylor and Francis, London, United Kingdom.

Jensen, F.V. et Nielsen, T. D. (2007). Bayesian networks and decision graphs. Statistics for Engineering and Information Science book series. Springer.

Naïm, P., Wuillemin, P.H., Leray, P., Pourret, O., and Becker, A. (2004). Les Réseaux Bayésiens. Eyrolles, Paris.

Pearl, J. (1988). Probabilistic reasoning in Intelligent Systems: Networks of plausible inference. Morgan Kaufman.

Pearl, J. (2003). Causality: Models, Reasoning, and Inference. *Econometric Theory*, **19**, 675-685.

Analyse de séries temporelles

Visualisation de séries temporelles

Utilisez cet outil pour créer en trois clics autant de graphiques que vous avez de séries temporelles.

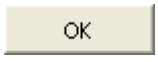
Dans cette section :

Description

Cet outil permet de créer en trois clics autant de graphiques que vous avez de séries temporelles. Il vous permet également de grouper les séries sur un même graphique. Enfin, une option vous permet de lier les graphiques aux données d'origine : si vous choisissez cette option, un changement des données entraîne alors automatiquement une mise à jour du graphique.

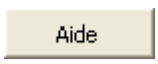
Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.



: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.

 : cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.

Onglet **Général**:

Séries temporelles : sélectionnez la ou les séries temporelles. Si des en- têtes de colonnes ont été sélectionnés, veuillez vérifier que l'option « Libellés des séries » est activée.

Données de date : activez cette option pour sélectionner des données de date. Ces données doivent être au format de data Excel, ou des valeurs numériques.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des séries : activez cette option si la première ligne des données sélectionnées (variables dépendantes et explicatives, libellés des observations, poids) contient un libellé.

Onglet **Graphiques**

Lier le graphique aux données d'entrée : activez cette option pour qu'une modification des données entraîne directement une modification des graphiques concernés.

Afficher toutes les séries sur un même graphique : activez cette option pour afficher toutes les séries sur un même graphique.

Résultats

Les graphiques sont affichés pour chacune des séries sélectionnées.

Exemple

Un exemple de visualisation de séries chronologiques est disponible sur le Centre d'aide XLSTAT. Pour accéder à cet exemple, veuillez vous connecter sur :

http://www.xlstat.com/demo-tsvizf.htm

Bibliographie

Brockwell P.J. and Davis R.A. (1996). Introduction to Time Series and Forecasting. Springer Verlag, New York.

Analyse descriptive

Utilisez cet outil pour calculer les statistiques descriptives adaptées aux séries chronologiques.

Dans cette section :

Description

L'une des phases essentielles de l'analyse des séries chronologiques consiste à déterminer si une valeur observée à un temps t dépend de ce qui a été observé dans le passé ou non. Si la réponse est affirmative, alors l'étape suivante essaiera de répondre à comment se manifeste cette dépendance.

Les fonctions d'autocovariance (FACV) et d'autocorrélation (FAC) estimées sur un échantillon donnent une idée de la dépendance entre les données d'une série. La visualisation de la FAC ou de la fonction d'autocorrélation partielle (FACP) aide à l'identification de modèles susceptibles de permettre d'expliquer un phénomène sur la base de ce qui a été observé, puis de prédire des valeurs futures. Par exemple la théorie montre que pour un modèle autorégressif d'ordre p, AR(p), la fonction FACP doit être nulle pour un décalage supérieur à p.

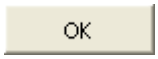
Les fonctions de corrélations croisées (FCC) permettent quant à elles de lier deux séries chronologiques et de déterminer si elles covarient, et si oui, dans quelle mesure.

Les fonctions FACV, FAC, FACP, FCC sont toutes calculées par cet outil.

Une autre étape importante de l'analyse des séries chronologiques consiste en la transformation des séries de manière à ne plus obtenir qu'un bruit blanc (voir Transformation de séries). Obtenir un bruit signifie que l'on a réussi à supprimer les autocorrélations, et les composants déterministes impliquant les variations de la série. Plusieurs tests sont proposés par XLSTAT pour tester, sur la base de la série elle même ou de sa FAC, si elle est significativement différente d'un bruit blanc ou non (Jarque Bera, Box-Pierce, Ljung-Box, McLeod-Li).

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.


: cliquez sur ce bouton pour changer la façon dont vous chargez les données dans XLSTAT. Si le bouton possède une icône de souris, XLSTAT vous permet de faire une sélection à la souris des données. Si le bouton possède une icône de liste, XLSTAT vous permet de sélectionner les données à partir d'une liste. Si le bouton possède une icône orange, des boutons complémentaires avec un point d'interrogation  sont affichés pour vous permettre d'importer les données à partir de fichiers.

Onglet **Général**:

Séries temporelles : sélectionnez la ou les séries temporelles que vous voulez analyser. Si des en-têtes de colonnes ont été sélectionnés, veuillez vérifier que l'option « Libellés des séries » est activée.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des séries : activez cette option si la première ligne des données sélectionnées (variables dépendantes et explicatives, libellés des observations, poids) contient un libellé.

Onglet **Options**:

Nombre de pas de temps : le nombre de pas de temps pour lesquels les statistiques sont calculées et affichées peut être soit déterminé de manière **automatique** par XLSTAT, soit fixé

par l'utilisateur.

Intervalles de confiance : activez cette option pour afficher des intervalles de confiance. La valeur que vous entrez (comprise entre 0.01 et 99.99) est utilisée pour déterminer les intervalles de confiance sur les estimations. Les intervalles de confiance sont automatiquement affichés sur les graphiques.

- **Hypothèse de bruit blanc** : activez cette option pour que les intervalles de confiance soient calculés sous hypothèse de bruit blanc.

Tests du bruit blanc : activez cette option pour que XLSTAT affiche les résultats concernant les tests du bruit blanc et le test de normalité de Jarque-Bera.

- **h1** : entrez le nombre minimum de pas de temps à prendre en compte pour le calcul des tests de bruit blanc.
- **h2** : entrez le nombre maximum de pas de temps à prendre en compte pour le calcul des tests de bruit blanc.
- **s** : entrez le nombre de pas de temps entre deux séries de tests. s doit être un multiple de (h2-h1).

Niveau de signification (%) : entrez le niveau de signification à utiliser pour les différents tests (valeur par défaut : 5%).

Onglet **Données manquantes** :

Ne pas accepter les valeurs manquantes : activez cette option pour que XLSTAT empêche la poursuite des calculs si des valeurs manquantes sont détectées.

Supprimer les observations : activez cette option pour supprimer les observations comportant des données manquantes.

Remplacer par la moyenne des valeurs précédente et suivante : activez cette option pour estimer les données manquantes par la moyenne de la première valeur précédente non manquante et de la première valeur suivante non manquante.

Ignorer les données manquantes : activez cette option pour ignorer les données manquantes.

Onglet **Sorties**:

Statistiques descriptives : activez cette option pour afficher les statistiques descriptives des séries sélectionnées.

Autocorrélations : activez cette option pour estimer la fonction d'autocorrélation des séries sélectionnées.

Autocovariances : activez cette option pour estimer la fonction d'autocovariance des séries sélectionnées.

Autocorrélations partielles : activez cette option pour calculer la fonction des autocorrélations partielles.

Corrélations croisées : activez cette option pour calculer la fonction des corrélations croisées.

Onglet **Graphiques** :

Autocorrélogramme : activez cette option pour afficher l'autocorrélogramme des séries sélectionnées.

Autocorrélogramme partiel : activez cette option pour afficher l'autocorrélogramme partiel des séries sélectionnées.

Corrélations croisées : activez cette option pour afficher le diagramme des corrélations croisées dans le cas où plusieurs séries ont été sélectionnées.

Résultats

Pour chaque série les résultats suivants sont affichés :

Statistiques simples : dans ce tableau sont affichés le nombre d'observations, le nombre d'observations manquantes, le minimum, le maximum, la moyenne, et l'écart-type non biaisé de la série.

Tests de normalité et de bruit blanc : dans ce tableau sont affichés les résultats des divers tests. Le test de normalité de Jarque-Bera est calculé une fois pour chacune des séries, alors que les tests du bruit blanc (Box- Pierce, Ljung-Box, Mcleod-Li) sont calculés pour chaque pas indiqué dans la boîte de dialogue. Le nombre de degrés de liberté (DDL), la valeur des statistiques et la p-value calculée sur la base d'une distribution du Khi²(DDL) sont affichés. Pour le test de Jarque-Bera, plus la p-value est faible, plus la normalité de l'échantillon degré est probable. Pour les trois autres tests, plus la p-value est faible, plus il est vraisemblable que les données correspondent à un bruit blanc.

Analyse descriptive : dans ce tableau sont affichés pour chaque pas de temps les valeurs des différentes fonctions descriptives, et les intervalles de confiance correspondants.

Graphiques : pour chaque fonction sélectionnée, un graphique est affiché si l'option correspondante a été activée dans la boîte de dialogue.

Si plusieurs séries ont été sélectionnées et que l'option « corrélations croisées » a été sélectionnée, les résultats suivants sont affichés :

Tests du bruit blanc : dans ce tableau sont affichés les résultats des tests de Box-Pierce, Ljung-Box, et Mcleod-Li, pour chaque nombre de pas de temps indiqué dans la boîte de dialogue. Le nombre de degrés de liberté (DDL), la valeur des statistiques et la p-value calculée sur la base d'une distribution du Khi²(DDL) sont affichés.

Corrélations croisées : dans ce tableau sont affichées pour chaque couple de variables les corrélations croisées. Le graphique correspondant est ensuite affiché.

Exemple

Un exemple d'analyse descriptive d'une série chronologique est disponible en permanence sur le Centre d'aide XLSTAT. Pour accéder à cet exemple, veuillez vous connecter sur :

http://www.xlstat.com/demo-descf.htm

Bibliographie

Box G. E. P. and Jenkins G. M. (1976). Time Series Analysis: Forecasting and Control. Holden-Day, San Francisco.

Box G. E. P. and Pierce D.A. (1970). Distribution of residual autocorrelations in autoregressive-integrated moving average time series models. *J Amer. Stat. Assoc.*, **65**, 1509-1526.

Brockwell P.J. and Davis R.A. (1996). Introduction to Time Series and Forecasting. Springer Verlag, New York.

Cryer, J. D. (1986). Time Series Analysis. Duxbury Press, Boston.

Fuller W.A. (1996). Introduction to Statistical Time Series, Second Edition. John Wiley & Sons, New York.

Jarque C.M. and Bera A.K. (1980). Efficient tests for normality, heteroscedasticity and serial independence of regression residuals. *Economic Letters*, **6**, 255-259.

Ljung G.M. and Box G. E. P. (1978). On a measure of lack of fit in time series models. *Biometrika*, **65**, 297-303.

McLeod A.I. and Li W.K. (1983). Diagnostic checking ARMA times series models using squares-residual autocorrelation. *J Time Series Anal.*, **4**, 269-273.

Shumway R.H. and Stoffer D.S. (2000). Time Series Analysis and Its Applications. Springer Verlag, New York.

Tests de Mann-Kendall

Utilisez cet outil pour déterminer avec un test non paramétrique si une tendance est identifiable dans une série temporelle, comprenant éventuellement une composante saisonnière.

Dans cette section :

Description

Ce test de tendance non paramétrique a d'abord été étudié par Mann (1945) puis repris par Kendall (1975) et amélioré par Hirsch (1982, 1984) qui a permis de prendre en compte une composante saisonnière.

L'hypothèse nulle H_0 de ces deux tests est qu'il n'y a pas de tendance. Les trois hypothèses alternatives de tendance négative, non nulle ou positive peuvent être choisies.

Les tests de Mann-Kendall s'appuient sur le calcul du tau de Kendall mesurant l'association entre deux échantillons et lui-même basé sur les rangs à l'intérieur des échantillons.

Test de tendance de Mann- Kendall

Dans le cas particulier du test de tendance, la première série est un indicateur temporel croissant généré automatiquement et pour lequel les rangs sont naturellement toujours croissants, ce qui simplifie les calculs. La statistique S de Kendall et sa variance sont données par

$$S = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \text{sgn}(x_j - x_i)$$

$$Var(S) = \frac{n(n-1)(2n+5)}{18}$$

Où n est le nombre de données de la série, et les x_i $(i = 1 \dots n)$ sont les observations, supposées indépendantes.

Pour le calcul de la p-value de ce test, XLSTAT permet de calculer, comme dans le cas du tau de Kendall, un test exact s'il n'y a pas d'ex-æquo dans les rangs des séries et si les tailles

d'échantillon sont inférieures à 50. Dans le cas où un calcul exact n'est pas possible, une approximation normale est utilisée, pour laquelle une correction de continuité est optionnelle mais recommandée.

Prise en compte de l'autocorrélation

Le test de tendance de Mann-Kendall requiert que les observations soient indépendantes, autrement dit la corrélation entre la série avec elle-même avec un décalage donné ne doit pas être significative. Dans le cas où il y a une autocorrélation dans la série, on montre que la variance peut être sous- estimée. Pour remédier à cela, plusieurs améliorations ont été suggérées. XLSTAT propose deux méthodes alternatives, la première publiée par Hamed et Rao (1998) et la seconde par Yue et Wang (2004). La première méthode fonctionne bien dans le cas d'absence de tendance dans la série, ce qui permet d'éviter de conclure qu'il y a une tendance lorsque cela est en fait due à l'autocorrélation. La seconde méthode présente l'avantage d'être plus fiable lorsqu'il y bien a une tendance et une autocorrélation. Les deux méthodes nécessitent le calcul de pente de Sen (Sen, P. K. (1968)). Cette valeur peut être affichée par XLSTAT si l'option correspondante est activée dans l'onglet sorties.

Il est évidemment recommandé, avant d'exécuter faire un test de tendance de Mann-Kendall, de vérifier d'abord les autocorrélations de la série étudiée en utilisant la fonction correspondante de XLSTAT-Time.

Test de Mann-Kendall avec saisonnalité

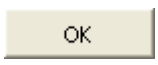
Dans le cas du test de Mann-Kendall avec saisonnalité, on tient compte du caractère saisonnier de la série. Autrement dit pour des données mensuelles ayant une saisonnalité de 12 mois, on ne va pas chercher à savoir s'il y une croissance au global sur la série, mais simplement si d'un mois de janvier à l'autre, d'un mois de février à l'autre, et ainsi de suite, il y une tendance.

Pour ce test, on calcule d'abord l'ensemble des tau de Kendall pour chaque saison, puis on calcule un tau de Kendall moyen. La variance de la statistique peut être calculée en faisant l'hypothèse que les séries sont indépendantes (par exemple les valeurs des mois de janvier et des mois de février sont indépendantes) ou dépendantes, ce qui requiert le calcul de covariances. XLSTAT permet les deux (dépendance sérielle ou non).

Pour le calcul de la p-value de ce test, XLSTAT utilise à une approximation normale pour la distribution de la moyenne des tau de Kendall Une correction de continuité peut être utilisée.

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

Annuler : cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

Aide : cliquez sur ce bouton pour afficher l'aide.

 : cliquez sur ce bouton pour rétablir les options par défaut.

 : cliquez sur ce bouton pour effacer les sélections de données.

Onglet **Général**:

Séries temporelles : sélectionnez la ou les séries temporelles que vous voulez analyser. Si des en-têtes de colonnes ont été sélectionnés, veuillez vérifier que l'option « Libellés des séries » est activée.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des séries : activez cette option si la première ligne des données sélectionnées (variables dépendantes et explicatives, libellés des observations, poids) contient un libellé.

Test de tendance de Mann-Kendall : activez cette option pour utiliser ce test.

Test de Mann-Kendall avec saisonnalité : activez cette option pour utiliser ce test. Entrez alors la valeur de la **période** (nombre de pas de temps entre deux mêmes saisons). Précisez également si vous considérez qu'il y a **dépendance sérielle** ou non.

Onglet **Options**:

Hypothèse alternative : choisissez l'hypothèse alternative à utiliser pour le test (voir la section description pour plus de détails).

Niveau de signification (%) : entrez la valeur du niveau de signification pour les tests (valeur par défaut : 5%).

p-values exactes : activez cette option si vous souhaitez que XLSTAT calcule la p-value exacte dans la mesure du possible (voir la section description pour plus de détails).

Correction de continuité : activez cette option si vous souhaitez que XLSTAT utilise la correction de continuité si le calcul d'une p-value exacte n'est pas demandé ou s'il n'est pas possible (voir la section description pour plus de détails).

Autocorrélations : activez l'une des deux options **Hamed et Rao** ou **Yue et Wang** pour prendre en compte d'éventuelles autocorrélations dans la série. Pour l'option Hamed et Rao vous avez la possibilité de filtrer les autocorrélations pour lesquelles la p-value est supérieure à un **seuil** que vous pouvez fixer (valeur par défaut : 10%).

Onglet **Données manquantes** :

Ne pas accepter les valeurs manquantes : activez cette option pour que XLSTAT empêche la poursuite des calculs si des valeurs manquantes sont détectées.

Supprimer les observations : activez cette option pour supprimer les observations comportant des données manquantes.

Remplacer par la moyenne des valeurs précédente et suivante : activez cette option pour estimer les données manquantes par la moyenne de la première valeur précédente non manquante et de la première valeur suivante non manquante.

Ignorer les données manquantes : activez cette option pour ignorer les données manquantes.

Onglet **Sorties**:

Statistiques descriptives : activez cette option pour afficher les statistiques descriptives des séries sélectionnées.

Pente de Sen : activez cette option pour afficher l'estimation de la pente de Sen. Vous pouvez également configurer la valeur de l'intervalle de confiance.

Résultats

Pour chaque série les résultats suivants sont affichés :

Statistiques simples : dans ce tableau sont affichés le nombre d'observations, le nombre d'observations manquantes, le minimum, le maximum, la moyenne, et l'écart-type non biaisé de la série.

Test de tendance de Mann-Kendall : les résultats du test de tendance Mann-Kendall sont affichés si l'option correspondante a été activée, suivis d'une aide à l'interprétation.

Test de Mann-Kendall avec saisonnalité : les résultats du test de Mann- Kendall avec saisonnalité sont affichés si l'option correspondante a été activée, suivis d'une aide à l'interprétation.

Exemple

Un exemple de test de tendance de Mann Kendall est disponible en permanence sur le Centre d'aide XLSTAT. Pour accéder à cet exemple, veuillez vous connecter sur :

http://www.xlstat.com/demo-mannkendallf.htm

Bibliographie

Hamed K.H. and Rao A.R. (1998). A modified Mann-Kendall trend test for autocorrelated data. *Journal of Hydrology*, **204** (1-4), 182-196.

Sen, P. K. (1968). Estimates of the regression coefficient based on Kendall's tau. Journal of the American Statistical Association, 63(324), 1379-1389.

Hirsch R.M., Slack, J.R., and Smith R.A. (1982). Techniques of trend analysis for monthly water quality data. *Water Resources Research*, **18**, 107-121.

Hirsch R.M. and Slack J.R. (1984). A nonparametric trend test for seasonal data with serial dependence. *Water Resources Research*, **20**, 727-732.

Kendall M. (1975). Multivariate Analysis. Charles Griffin & Company, London.

Mann H.B. (1945). Nonparametric tests against trend. *Econometrica*, **13**, 245-259.

Yue S and Wang C.Y. (2004). The Mann-Kendall test modified by effective sample size to detect trend in serially correlated hydrological series. *Water Resour. Manag.*, **18**, 201-218.

Tests d'homogénéité

Utilisez cet outil pour déterminer avec l'un des quatre tests proposés (Pettitt, Buishand, SNHT ou von Neumann), si on peut considérer qu'une série est homogène dans le temps, ou s'il existe un temps auquel se produit un décalage de la série.

Dans cette section :

Description

Les tests d'homogénéité rassemblent un grand nombre de tests pour lesquels l'hypothèse nulle est qu'une série temporelle est homogène entre deux temps donnés.

La variété des tests vient de ce que les hypothèses alternatives possibles sont nombreuses : changement de distribution, changements de moyenne (une ou plusieurs fois) ou présence de tendance.

Les tests présentés dans cet outil correspondent à l'hypothèse alternative d'un unique décalage. Pour l'ensemble des tests, XLSTAT fournit des p-values en utilisant des rééchantillonnages Monte Carlo, les calculs exacts étant soit impossibles soit trop coûteux en temps de calcul.

Pour la présentation des différents tests, nous désignons par $X_i(i = 1, 2, \ldots, T)$ une série de T variables dont on observe une valeur $x_i(i = 1, 2, 3, \ldots, T)$ à T temps successifs. Soit $\hat{\mu}$ la moyenne des T valeurs observées, et soit $\hat{\sigma}$ leur écart type biaisé (on divise par T).

Remarque 1 : si l'on a une idée précise du temps de changement, les tests déjà existant dans la section des tests paramétriques ou non paramétriques peuvent être utilisés : par exemple si l'on suppose que les variables suivent des distributions normales, on peut utiliser le test z (variance connue) ou de Student (variance estimée) pour tester la présence d'un changement de moyenne à un temps t. Si l'on pense que la variance change, on peut utiliser un test de comparaison de variances (test de Fisher dans le cas normal par exemple, ou de Kolmogorov-Smirnov dans un cas plus général).

Remarque 2 : les tests présentés ci-dessous sont sensibles à une tendance (linéaire par exemple). Avant d'appliquer ces tests, il faut donc bien être sûr de vouloir identifier un temps de rupture séparant deux périodes homogènes.

Test de Pettitt

Le test de Pettitt est un test non paramétrique ne nécessitant aucune hypothèse quant à la distribution des données. Le test de Pettitt est une adaptation du test de Mann-Whitney basé sur les rangs, permettant d'identifier le temps auquel se produit un changement.

Dans son article de 1979 Pettitt décrit l'hypothèse nulle comme étant que les T variables suivent une même distribution F, et l'hypothèse alternative comme étant qu'à un temps τ se produit un changement de distribution. Néanmoins le test de Pettitt ne permet pas de détecter un changement de distribution s'il n'est pas assorti d'un changement de position. Par exemple, si avant le temps τ, les variables suivent une distribution normale $N(0,1)$ et à partir du temps τ une distribution $N(0,3)$, le test de Pettitt ne détectera pas de changement, de la même manière qu'un test de Mann-Whitney ne permettrait pas de détecter un changement de position dans un tel cas. Il faudrait dans ce cas utiliser par une méthode s'appuyant exemple sur le test de Kolmogorov Smirnov. Nous reformulons donc ainsi les hypothèses nulle et alternatives :

- H0 : les T variables suivent une ou plusieurs distributions ayant un même paramètre de position.
- Test bilatéral : Ha : il existe un temps τ à partir duquel les variables changent de paramètre de position.
- Test unilatéral à gauche : Ha : il existe un temps τ à partir duquel le paramètre de position des variables diminue de Δ.
- Test unilatéral à droite : Ha : il existe un temps τ à partir duquel le paramètre de position des variables augmente de Δ.

La statistique du test de Pettitt est calculée comme suit :

On pose $D_{ij} = -1$ si $(x_i - x_j) < 0$, $D_{ij} = 0$ si $(x_i - x_j) = 0$, $D_{ij} = 1$ si $(x_i - x_j) > 0$

On définit ensuite $U_{t,T} = \sum_{i=1}^{t} \sum_{j=i+1}^{T} D_{ij}$

La statistique de Petitt correspondant à chacune des hypothèses alternatives est définie par :

$K_T = \max_{1 \leq t < T} |U_{t,T}|$, pour le test bilatéral

$K_T^+ = \max_{1 \leq t < T} U_{t,T}$, pour le test unilatéral à gauche

$K_T^- = -\min_{1 \leq t < T} U_{t,T}$, pour le test unilatéral à droite

XLSTAT évalue la p-value ainsi qu'un intervalle autour de la p-value en calculant par rééchantillonnage les valeurs possible de la statistique K.

Test SNHT d'Alexandersson

Le test SNHT (*Standard normal homogeneity test*) a été développé par Alexandersson (1986) pour détecter un changement dans une série de précipitations. Le test s'applique à une série de ratios comparant les observations d'une station de mesure à la moyenne de plusieurs stations. Les ratios sont ensuite centrés-réduits. La série des X_i correspond ici aux ratios standardisés. Les hypothèses nulle et alternative sont définies par :

- H0 : les T variables X_i suivent une loi $N(0,1)$.
- Ha : entre les temps 1 et ν les variables sont distribuées suivant une loi $N(\mu_1,1)$ et entre $\nu+1$ et T elles sont distribuées suivant une loi $N(\mu_2,1)$.

La statistique d'Alexandersson est définie par :

$$T_0 = \max_{1\leq t<T}[\nu\overline{z}_1^2 + (n-\nu)\overline{z}_2^2]$$

avec

$$\begin{aligned}\overline{z}_1 &= \frac{1}{\nu}\sum_{t=1}^{\nu} x_t \\ \overline{z}_2 &= \frac{1}{n-\nu}\sum_{t=\nu+1}^{T} x_i\end{aligned}$$

La statistique To dérive d'un calcul comparant la vraisemblance des deux modèles alternatifs. Le modèle correspondant à Ha implique que l'on estime μ_1 et μ_2 tout en déterminant le paramètre ν maximisant la vraisemblance.

XLSTAT évalue la p-value ainsi qu'un intervalle autour de la p-value en calculant par simulation les valeurs possible de la statistique To.

Remarque : si ν est connu, il suffit de faire un test z sur les deux séries de ratio. Le test SNHT permet de déterminer le ν le plus probable.

Test de Buishand

Le test de Buishand (1982) peut être utilisé sur des variables suivant des distributions quelconques. Néanmoins ses propriétés ont été particulièrement étudiées pour le cas normal. L'article Buishand se concentre sur le cas du test bilatéral, mais pour la statistique Q présentée ci-dessous le cas unilatéral est aussi possible. Buishand a développé une seconde statistique R, pour laquelle seule une hypothèse bilatérale est possible.

Dans le cas de la statistique Q, les hypothèses nulle et alternative sont définies par :

- H0 : les T variables suivent une ou plusieurs distributions ayant une même moyenne.
- Test bilatéral : Ha : Il existe un temps τ à partir duquel les variables changent de moyenne.

- Test unilatéral à gauche : Ha : Il existe un temps τ à partir duquel la moyenne des variables diminue de Δ.
- Test unilatéral à droite : Ha : Il existe un temps τ à partir duquel la moyenne des variables augmente de Δ.

On définit $S_o^* = 0$, $S_k^* = \sum_{i=1}^{k}(x_i - \hat{\mu})$,$k = 1, 2, \ldots, T$ et $S_k^{**} = \frac{S_k^*}{\hat{\sigma}}$

La statistique du test du Q de Buishand est calculée comme suit :

$Q = \max_{1 \leq k < T} |S_k^{**}|$, pour le test bilatéral

$Q^- = \max_{1 \leq k < T}(S_k^{**})$, pour le test unilatéral à gauche

$Q^+ = -\min_{1 \leq k < T}(S_k^{**})$, pour le test unilatéral à droite

XLSTAT évalue la p-value ainsi qu'un intervalle autour de la p-value en calculant par rééchantillonnage les valeurs possible de la statistique Q.

Dans le cas de la statistique R (R pour *Range*, l'amplitude), les hypothèses nulle et alternative sont définies par :

- H0 : les T variables suivent une ou plusieurs distributions ayant une même moyenne.
- Test bilatéral : Ha : Les T variables ne sont pas homogènes en moyenne.

La statistique du test du R de Buishand est calculée comme suit :

$$R = \max_{1 \leq k < T}(S_k^{**}) - \min_{1 \leq k < T}(S_k^{**})$$

XLSTAT évalue la p-value ainsi qu'un intervalle autour de la p-value en calculant par simulation les valeurs possible de la statistique R.

Remarque : le test de R ne permet pas de déterminer le temps de rupture.

Test du rapport de von Neumann

Le rapport de von Neumann est définit par :

$$N = \frac{1}{T\hat{\sigma}} \sum_{i=1}^{T-1}(x_i - x_{i+1})^2$$

On montre que l'espérance de N est 2 lorsque dans le cas où les X_i sont de même moyenne.

XLSTAT évalue la p-value ainsi qu'un intervalle autour de la p-value en calculant par simulation les valeurs possible de la statistique N.

Remarque : le test de N ne permet pas de déterminer le temps de rupture.

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.

OK : cliquez sur ce bouton pour lancer les calculs.

Annuler : cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

Aide : cliquez sur ce bouton pour afficher l'aide.

 : cliquez sur ce bouton pour rétablir les options par défaut.

 : cliquez sur ce bouton pour effacer les sélections de données.

Onglet **Général**:

Séries temporelles : sélectionnez la ou les séries temporelles que vous voulez analyser. Si des en-têtes de colonnes ont été sélectionnés, veuillez vérifier que l'option « Libellés des séries » est activée.

Données de date : activez cette option pour sélectionner des données de date. Ces données doivent être au format de data Excel, ou des valeurs numériques.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des séries : activez cette option si la première ligne des données sélectionnées (variables dépendantes et explicatives, libellés des observations, poids) contient un libellé.

Test de Pettitt : activez cette option pour utiliser ce test (voir la section description pour plus de détails).

Test SNHT : activez cette option pour utiliser ce test (voir la section description pour plus de détails).

Test de Buishand : activez cette option pour utiliser ce test (voir la section description pour plus de détails).

Test de von Neumann : activez cette option pour utiliser ce test (voir la section description pour plus de détails).

Onglet **Options**:

Hypothèse alternative : choisissez l'hypothèse alternative à utiliser pour le test (voir la section description pour plus de détails).

Niveau de signification (%) : entrez la valeur du niveau de signification pour les tests (valeur par défaut : 5%).

Méthode Monte Carlo : activez cette option pour calculer la p-value en utilisant des simulations Monte Carlo. Entrez alors le nombre maximum de simulations à réaliser et le temps de calcul maximum à ne pas dépasser, exprimé en secondes.

Onglet **Données manquantes**:

Ne pas accepter les valeurs manquantes : activez cette option pour que XLSTAT empêche la poursuite des calculs si des valeurs manquantes sont détectées.

Supprimer les observations : activez cette option pour supprimer les observations comportant des données manquantes.

Remplacer par la moyenne des valeurs précédente et suivante : activez cette option pour estimer les données manquantes par la moyenne de la première valeur précédente non manquante et de la première valeur suivante non manquante.

Onglet **Sorties**:

Statistiques descriptives : activez cette option pour afficher les statistiques descriptives des séries sélectionnées.

Onglet **Graphiques**:

Afficher les graphiques : activez cette option pour afficher les graphiques permettant de comparer les séries avant et après transformation.

Résultats

Pour chaque série les résultats suivants sont affichés :

Statistiques simples : dans ce tableau sont affichés le nombre d'observations, le nombre d'observations manquantes, le minimum, le maximum, la moyenne, et l'écart-type non biaisé de la série.

Les résultats des différents tests sont ensuite affichés. Pour les tests de Pettitt, le SNHT et le test du Q de Buishand des graphiques sont affichés avec les moyennes μ_1 et μ_2 si une rupture est détectée et μ si aucune rupture n'est détectée.

Exemple

Un exemple de test d'homogénéité est disponible sur le Centre d'aide XLSTAT. Pour accéder à cet exemple, veuillez vous connecter sur :

http://www.xlstat.com/demo-homogeneityf.htm

Bibliographie

Alexandersson H. (1986). A homogeneity test applied to precipitation data. *Journal of Climatology* , **6**, 661-675.

Buishand T.A. (1982). Some methods for testing the homogeneity of rainfall data. *Journal of Hydrology*, **58**, 11-27.

Pettitt A.N. (1979). A non-parametric approach to the change-point problem. *Appl. Statist.*, **28(2)**, 126-135.

Von Neumann J. (1941). Distribution of the ratio of the mean square successive difference to the variance. *Ann. Math. Stat.*, **12**, :367-395.

Test de Durbin-Watson

Utilisez ce module pour tester la présence d'autocorrélation dans les résidus d'une régression linéaire.

Dans cette section :

Description

Développé par J.Durbin et G.Watson (1950, 1951), le test de Durbin-Watson est utilisé pour détecter l'autocorrélation entre les résidus d'une régression linéaire.

Test de Durbin-Watson

Soit Y la variable dépendante, X la matrice des variables exploratoires, α et β les coefficients et ϵ le terme d'erreur. On considère le modèle suivant :

$$y_t = \alpha + \beta x_t + \epsilon_t$$

Dans la pratique, les termes d'erreurs sont parfois autocorrélés, ce qui peut entraîner une mauvaise estimation des paramètres. Le test de Durbin-Watson est utilisé dans le but de détecter ces autocorrélations.

On suppose que les $\{\epsilon_t\}$ sont stationnaires et distribués selon une loi normale de moyenne 0. Les hypothèses nulle et alternative du test de Durbin-Watson sont les suivantes :

- H0: Les résidus ne sont pas autocorrélés
- Ha: Les résidus sont distribués selon un AR(p)

$AR(p)$ désigne un processus autorégressif d'ordre p.

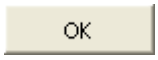
La statistique D du test s'écrit :

$$D = \frac{\sum_{t=p+1}^{n} (\epsilon_t - \epsilon_{t-p})^2}{\sum_{t=r+1}^{n} \epsilon_t^2}$$

Dans le cadre du test de Durbin-Watson, le principal problème réside dans l'estimation des p-values associées au test. En effet, il n'existe pas de formule explicite. XLSTAT utilise l'algorithme de Pan pour les séries de moins de 70 observations et la procédure proposée par Imhof (1961) pour celles ayant plus de 70 observations.

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.

Onglet **Général**:

Résidus : Sélectionnez les résidus issus de la régression linéaire. Si le libellé des variables a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

X / Variables explicatives : sélectionnez la ou les variables quantitatives explicatives sur la feuille Excel. Les données sélectionnées doivent être de type numérique. Si le libellé des variables a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des variables : activez cette option si la première ligne des données sélectionnées (variables dépendantes et explicatives, libellés des observations, poids) contient un libellé.

Onglet **Options**:

Niveau de signification (%) : entrez le niveau de signification à utiliser pour le test (valeur par défaut : 5%)

Ordre : entrez l'ordre du test, c'est-à-dire le nombre de retards r pour les résidus (valeur par défaut : 1)

Onglet **Données manquantes** :

Ne pas accepter les valeurs manquantes : activez cette option pour que XLSTAT empêche la poursuite des calculs si des valeurs manquantes sont détectées.

Supprimer les observations : activez cette option pour supprimer les observations comportant des données manquantes.

Remplacer par la moyenne des valeurs précédente et suivante : activez cette option pour estimer les données manquantes par la moyenne de la première valeur précédente non manquante et de la première valeur suivante non manquante.

Ignorer les données manquantes : activez cette option pour ignorer les données manquantes.

Onglet **Sorties**:

Statistiques descriptives : activez cette option pour afficher les statistiques descriptives des séries sélectionnées.

Résultats

Statistiques descriptives : le tableau de statistiques descriptives présente des statistiques simples pour les résidus. Les statistiques telles que le nombre d'observations, le nombre de données manquantes, le nombre de données non manquantes, la moyenne, et l'écart-type (non biaisé) sont affichés.

Les résultats du test de Durbin-Watson sont ensuite affichés.

Exemple

Un exemple d'utilisation du test de Durbin-Watson est disponible sur le Centre d'aide XLSTAT à l'adresse :

http://www.xlstat.com/demo-durbinwatsonf.htm

Bibliographie

Durbin J. and Watson G. S. (1950). Testing for serial correlation in least squares regression, I. *Biometrika*, **37(3-4)**, 409-428.

Durbin J. and Watson G. S. (1951). Testing for serial correlation in least squares regression, II. *Biometrika*, **38(1-2)**, 159-179.

Farebrother R. W. (1980). Algorithm AS 153. Pan's procedure for the tail probabilities of the Durbin–Watson statistic. *Applied Statistics*, **29**, 224-227.

Imhof J.P. (1961), Computing the Distribution of Quadratic Forms of Normal Variables. *Biometrika*, **48**, 419-426.

Kim M. (1996). A remark on algorithm AS 279: computing p-values for the generalized Durbin-Watson statistic and residual autocorrelation in regression. *Applied Statistics*, **45**, 273-274

Kohn R., Shively T. S. and Ansley C. F. (1993). Algorithm AS 279: Computing p-values for the generalized Durbin-Watson statistic and residual autocorrelations in regression. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, **42(1)**, 249-258

Pan J.-J. (1968). Distribution of noncircular correlation coefficients. *Selected Transactions in Mathematical Statistics and Probability*, 281-291.

Estimation de Cochrane-Orcutt

Utilisez cette fonction pour prendre en compte une corrélation de type AR(1) dans le terme d'erreur du modèle linéaire.

Dans cette section :

Description

Développée par D.Cochrane et G. Orcutt en 1949, l'estimation de Cochrane- Orcutt est une méthode très employée en économétrie qui permet de prendre en compte une corrélation d'ordre 1 (AR(1)) pour le terme d'erreur du modèle linéaire classique. En effet, dans le cas où le terme d'erreur est généré selon un processus AR(1), les méthodes d'estimation usuelles ne sont pas adaptées car les écarts-type sont estimés avec un biais.

Estimation de Cochrane-Orcutt

Soit Y une variable dépendante, et X la matrice des variables explicatives, α et β les coefficients et ϵ le terme d'erreur. On considère le modèle suivant :

$$y_t = \alpha + \beta x_t + \epsilon_t$$

On suppose que le terme d'erreur ϵ est généré selon un processus stationnaire autorégressif d'ordre 1 tel que :

$$\epsilon_t = \rho\epsilon_{t-1} + \epsilon_t, \text{ mit } |\rho| < 1$$

où ϵ_t est un bruit blanc.

Afin d'estimer les paramètres du modèle, la méthode de Cochrane-Orcutt se base sur le modèle suivant :

$$\forall t \geq 2, y_t - \rho y_{t-1} = \alpha(1-\rho) + \beta(X_t - \rho X_{t-1}) + \epsilon_t$$

En introduisant les trois nouvelles variables suivantes

$$Y^* = y_t - \rho y_{t-1}, X^* = X_t - \rho X_{t-1}, \lambda^* = 1 - \rho$$

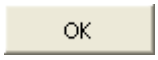
le modèle peut être réécrit sous la forme :

$$\forall t \geq 2, y_t^* = \alpha\lambda^* + \beta X_t^* + e_t$$

Etant donné que $\{e_t\}$ est un bruit blanc, la méthode d'inférence usuelle peut être utilisée.

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.

Onglet **Général**:

Y / Variables dépendantes :

Quantitatives : sélectionnez la ou les variables réponses que vous souhaitez modéliser. Si plusieurs variables sont sélectionnées, XLSTAT fera les calculs pour chacune des variables indépendamment. Si des en-têtes de colonnes ont été sélectionnés, veuillez vérifier que l'option « Libellés des variables » est activée.

X / Variables explicatives :

Quantitatives : sélectionnez la ou les variables quantitatives explicatives sur la feuille Excel. Les données sélectionnées doivent être de type numérique. Si le libellé des variables a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

Données de dates : sélectionnez les données correspondantes aux dates relevées. Ces données doivent être numériques. Le nombre de lignes doit être égal au nombre de colonnes du tableau précédent. Si un en-tête a été sélectionné sur la première ligne, veillez à ce que l'option «libellés des colonnes» soit activée.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des variables : activez cette option si la première ligne des données sélectionnées (variables dépendantes et explicatives, libellés des observations, poids) contient un libellé.

Libellés des observations : activez cette option si vous voulez utiliser des libellés d'observations pour l'affichage des résultats. Si l'option « Libellés des variables » est activée, la première cellule de la sélection doit comprendre un en-tête. Si vous n'activez pas cette option, des libellés seront automatiquement créés (Obs1, Obs2, …).

Poids des observations : activez cette option si vous voulez pondérer les observations. Si vous n'activez pas cette option, les poids seront tous considérés comme valant 1. Les poids doivent être impérativement supérieurs ou égaux à 0. Un poids de 2 est équivalent à répéter deux fois la même observation. Si un en-tête de colonne a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

Poids dans la régression : activez cette option si vous voulez effectuer une régression par les moindres carrés pondérés. Si vous n'activez pas cette option, les poids seront tous considérés comme valant 1. Les poids doivent être impérativement supérieurs ou égaux à 0. Si un en-tête de colonne a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

Onglet **Options**:

Tolérance : activez cette option pour permettre à l'algorithme de calcul de la régression OLS ne pas prendre en compte les variables qui seraient soit constantes soit trop corrélées avec d'autres variables déjà utilisées dans le modèle (0.0001 par défaut).

Intervalle de confiance (%) : entrez l'étendue en pourcentage de l'intervalle de confiance à utiliser pour les différents tests, et pour le calcul des intervalles de confiance autour des paramètres et des prédictions. Valeur par défaut : 95.

Onglet **Validation** :

Validation : activez cette option si vous souhaitez utiliser une partie des données sélectionnées pour valider le modèle.

Jeu de validation : choisissez l'une des options pour définir le mode de sélection des observations utilisées pour la validation :

- **Aléatoire** : les observations sont sélectionnées de manière aléatoire. Le « Nombre d'observations » doit alors être saisi.
- **N dernières lignes** : les N dernières observations sont sélectionnées pour la validation. Le « Nombre d'observations » N doit alors être saisi.
- **N premières lignes** : les N premières observations sont sélectionnées pour la validation. Le « Nombre d'observations » N doit alors être saisi.
- **Variable de groupe** : si vous choisissez cette option, vous devez ensuite sélectionner une variable indicatrice composée de 0 pour les observations à utiliser pour le calcul du modèle, et de 1 pour les observations à utiliser pour la validation du modèle.

Onglet **Prédiction** :

Prédiction : activez cette option si vous souhaitez sélectionner des données à utiliser en mode prédiction. Si vous activez cette option, vous devez veiller à ce que les données de prédiction soient organisées comme les données d'estimation : mêmes variables, même ordre dans les sélections. En revanche vous ne devez pas sélectionner de libellés de variables : la première ligne des sélections décrites ci-dessous doit être une ligne de données.

X / Variables explicatives : sélectionner la ou les variables quantitatives explicatives. La première ligne ne doit pas comprendre d'en- tête.

Libellés des observations : activez cette option si vous voulez utiliser des libellés d'observations disponibles sur une feuille Excel pour l'affichage des résultats. La première ligne ne doit pas comprendre d'en-tête. Si vous n'activez pas cette option, des libellés seront automatiquement créés (PredObs1, PredObs2, …).

Onglet **Données manquantes** :

Supprimer les observations : activez cette option pour supprimer les observations comportant des données manquantes.

- **Vérifier pour chaque Y séparément** : choisissez cette option si vous voulez que lorsque, pour une observation donnée, il y a des données manquantes uniquement dans les Y, l'observation ne soit supprimée que si la donnée correspondante au Y en cours de modélisation est manquante.
- **Pour tous les Y** : choisissez cette option pour supprimer toutes les observations pour lesquelles des Y sont manquants.
- Remarque : les deux alternatives ci-dessus sont sans effet si il n'y a qu'un seul Y.

Estimer les données manquantes : activez cette option pour estimer les données manquantes avant le début des calculs.

- **Moyenne ou mode** : activez cette option pour estimer les données manquantes en utilisant la moyenne (variables quantitatives) ou le mode (variables qualitatives) pour les variables correspondantes.
- **Plus proche voisin** : activez cette option pour estimer les données manquantes d'une observation en recherchant le plus proche voisin de l'observation.

Onglet **Sorties**:

Statistiques descriptives : activez cette option pour afficher les statistiques descriptives pour les variables sélectionnées.

Corrélations : activez cette option pour afficher la matrice de corrélation pour les variables quantitatives (dépendantes et explicatives).

Analyse de la variance : activez cette option pour afficher le tableau d'analyse de la variance.

Coefficients normalisés : activez cette option pour afficher les paramètres normalisés du modèle (coefficients bêta).

Prédictions et résidus : activez cette option pour afficher les prédictions et les résidus pour l'ensemble des observations.

Onglet **Graphiques** :

Graphiques de régression : activez cette option pour afficher les graphiques de régression :

- **Coefficients normalisés** : activez cette option pour afficher sur un graphique les paramètres normalisés du modèle avec leur intervalle de confiance.
- **Prédictions et résidus** : activez cette option pour afficher les graphiques suivants :

(1) Droite de régression : ce graphique n'est affiché que s'il n'y a qu'une seule variable explicative, et que cette variable est quantitative.

(2) Variable explicative versus résidus normalisés : ce graphique n'est affiché que s'il n'y a qu'une seule variable explicative, et que cette variable est quantitative.

(3) Variable dépendante versus résidus normalisés.

(4) Prédictions pour la variable dépendante versus variable dépendante.

(5) Graphique en bâtons des résidus normalisés.

- **Intervalles de confiance** : activez cette option pour afficher les intervalles de confiance sur les graphiques (1) et (4).

Résultats

Statistiques descriptives : les tableaux de statistiques descriptives présentent pour toutes les variables sélectionnées des statistiques simples. Pour les variables dépendantes (en bleu), les variables explicatives et les variables instrumentales, sont affichés le nombre d'observations, le nombre de données manquantes, le nombre de données non manquantes, la moyenne, et l'écart-type (non biaisé).

Matrice de corrélation : ce tableau est affiché afin de vous permettre d'avoir un aperçu des corrélations entre les différentes variables sélectionnées.

Coefficients d'ajustement : dans ce tableau sont affichées les statistiques relatives à l'ajustement du modèle de régression :

- **Observations** : le nombre d'observations prises en compte dans les calculs. Dans les formules présentées ci-dessous n désigne le nombre d'observations.
- **Somme des poids** : la somme des poids des observations prises en compte dans les calculs. Dans les formules présentées ci-dessous W désigne la somme des poids.
- **DDL** : le nombre de degrés de liberté pour le modèle retenu (correspondant à la partie erreurs).
- **R²** : le coefficient de détermination du modèle. Ce coefficient, dont la valeur est comprise entre 0 et 1, n'est affiché que si la constante du modèle n'est pas fixée par l'utilisateur. Sa valeur est définie par
- **R²** : le coefficient de détermination du modèle. Sa valeur est comprise entre 0 et 1. Il est défini par :

$$R^2 = \frac{\sum_{i=1}^{n} w_i(\hat{y}_i - \bar{y}_i)^2}{\sum_{i=1}^{n} w_i(y_i - \bar{y}_i)^2} \text{ avec } \bar{y} = \frac{1}{W}\sum_{i=1}^{n} w_i y_i$$

Le R^2 s'interprète comme la proportion de la variabilité de la variable dépendante expliquée par le modèle. Plus le R^2 est proche de 1, meilleur est le modèle. L'inconvénient du R^2 est qu'il ne prend pas en compte le nombre de variables utilisées pour ajuster le modèle.

- **R² ajusté** : le coefficient de détermination ajusté du modèle. Le R^2 ajusté peut être négatif si le R^2 est voisin de zéro. Ce coefficient n'est affiché que si la constante du modèle n'est pas fixée par l'utilisateur. Il est défini par :

$$\hat{R}^2 = 1 - (1 - R^2)\frac{W - 1}{W - p - 1}$$

Le R^2 ajusté est une correction du R^2 qui permet de prendre en compte le nombre de variables utilisées dans le modèle.

- **MCE** : la moyenne des carrés des erreurs (MCE) est définie par :

$$MCE = \frac{1}{W - p^*}\sum_{i=1}^{n} w_i(y_i - \hat{y}_i)^2$$

- **RMCE** : la racine de la moyenne des carrés des erreurs (RMCE) est la racine carrée de la MCE.

- **MAPE** : la *Mean Absolute Percentage Error* est calculée comme suit :

$$MAPE = \frac{100}{W}\sum_{i=1}^{n} w_i \left| \frac{y_i - \hat{y}_i}{y_i} \right|$$

- **DW** : le coefficient de Durbin-Watson est défini par

$$DW = \frac{\sum_{i=2}^{n}[(y_i - \hat{y}_i) - (y_{i-1} - \hat{y}_{i-1})]^2}{\sum_{i=1}^{n} w_i (y_i - \hat{y}_i)^2}$$

Ce coefficient correspond au coefficient d'autocorrélation d'ordre 1 et permet de vérifier que les résidus du modèle ne sont pas autocorrélés, sachant que l'indépendance des résidus est l'une des hypothèses de base de la régression linéaire. L'utilisateur pourra se référer à une table des coefficients de Durbin-Watson pour vérifier si l'hypothèse d'indépendance des résidus est acceptable.

- **Cp** : le coefficient Cp de Mallows est défini par

$$Cp = \frac{SCE}{\hat{\sigma}} + 2p^* - W$$

où SCE est la somme du carré des erreurs pour le modèle avec p variables explicatives, et où $\hat{\sigma}$ correspond à l'estimateur de la variance des résidus pour le modèle comprenant toutes les variables explicatives. Plus le coefficient Cp est proche de p* moins le modèle est biaisé.

- **AIC** : le critère d'information d'Akaike (Akaike's Information Criterion) est défini par

$$AIC = W \ln(\frac{SCE}{W}) + 2p^*$$

Ce critère proposé par Akaike (1973) dérive de la théorie de l'information, et s'appuie sur la mesure de Kullback et Leibler (1951). C'est un critère de sélection de modèles qui pénalise les modèles pour lesquels l'ajout de nouvelles variables explicatives n'apporte pas suffisamment d'information au modèle, l'information étant mesurée au travers de la SCE. On cherche à minimiser le critère AIC.

- **SBC** : le critère bayésien de Schwarz (Schwarz's Bayesian Criterion) est défini par

$$SBC = W \ln(\frac{SCE}{W}) + \ln(W)p^*$$

Ce critère proposé par Schwarz (1978) est proche du critère AIC, et comme ce dernier on cherche à le minimiser.

- **PC** : le critère de prédiction d'Amemiya (Amemiya's Prediction Criterion) est défini par

$$PC = \frac{(1 - R^2)(W + p^*)}{W - p^*}$$

Ce critère proposé par Amemiya (1980) permet comme le R² ajusté de tenir compte de la parcimonie du modèle.

- **Press** : la statistique du Press (predicted residual error sum of squares) n'est affichée que si l'option correspondante a été activée dans la boîte de dialogue. Elle est définie par

$$Press = \sum_{i=1}^{n} w_i (y_i - \hat{y}_{i(-i)})^2$$

où $\hat{y}_{i(-i)}$ est la prédiction pour l'observation i lorsque cette dernière n'est pas utilisée pour l'estimation des paramètres. On obtient alors

$$\text{Press RMCE } = \sqrt{\frac{\text{Press}}{W - p^*}}$$

Le Press RMCE peut alors être comparé au RMCE. Une différence importante entre les deux indique que le modèle est sensible à la présence ou absence de certaines observations dans le modèle.

Le **tableau d'analyse de la variance** permet d'évaluer le pouvoir explicatif des variables explicatives. Dans le cas où la constante du modèle n'est pas fixée à une valeur donnée, le pouvoir explicatif est évalué en comparant l'ajustement (au sens des moindres carrés) du modèle final avec l'ajustement du modèle rudimentaire composé d'une constante égale à la moyenne de la variable dépendante. Dans le cas où la constante du modèle est fixée, la comparaison est faite par rapport au modèle pour lequel la variable dépendante serait égale à la constante fixée.

L'**équation du modèle** est ensuite affichée pour faciliter la lecture ou la réutilisation du modèle.

Le tableau des **coefficients normalisés** (aussi appelés coefficients bêta) permet de comparer le poids relatif des variables. Plus la valeur absolue d'un coefficient est élevée, plus le poids de la variable correspondante est important. Lorsque l'intervalle de confiance autour des coefficients normalisés comprend la valeur 0 (cela est facilement visible sur le graphique des coefficients normalisés), le poids d'une variable dans le modèle n'est pas significatif. Le coefficient d'autocorrélationρest également disponible.

Dans le tableau des **prédictions et résidus** sont donnés pour chaque observation, son poids, la valeur de la variable explicative qualitative s'il n'y en a qu'une, la valeur observée de la variable dépendante, la prédiction du modèle, les résidus, les intervalles de confiance et la prédiction ajustée. Deux types d'intervalles de confiance sont affichés : un intervalle de confiance autour de la moyenne (correspondant au cas où l'on ferait la prédiction pour un nombre infini d'observations avec un ensemble de valeurs données des variables explicatives) et un intervalle autour de la prédiction ponctuelle (correspondant au cas d'une prédiction isolée pour des valeurs données des variables explicatives). Le second intervalle est toujours plus grand que le premier, les aléas étant plus importants. Si des données de validation ont été sélectionnées, elles sont affichées en fin de tableau.

Les **graphiques** qui suivent permettent de visualiser les résultats mentionnés ci-dessus. S'il n'y a qu'une seule variable explicative dans le modèle, le premier graphique affiché permet de visualiser les valeurs observées, la droite de régression et les deux types d'intervalles de confiance autour des prévisions. Le second graphique permet quant à lui de visualiser les résidus normalisés en fonction de la variable explicative. En principe, les résidus doivent être distribués de manière aléatoire autour de l'axe des abscisses. L'observation d'une tendance ou d'une forme révèlerait un problème au niveau du modèle.

Les **trois graphiques** affichés ensuite permettent de visualiser respectivement l'évolution des résidus normalisés en fonction de la variable dépendante, la distance entre les prédictions et les observations (pour un modèle idéal, les points seraient tous sur la bissectrice), et les résidus normalisés sur la forme d'un diagramme en bâtons. Ce dernier graphique permet de rapidement voir si un nombre anormal de données sort de l'intervalle]-2, 2[sachant que ce dernier, sous hypothèse de normalité, doit contenir environ 95% des données.

Si vous avez sélectionné des données à utiliser pour calculer des **prédictions sur de nouvelles observations**, le tableau correspondant est ensuite affiché.

Exemple

Un exemple d'utilisation de la méthode de Cochrane Orcutt est disponible sur le Centre d'aide XLSTAT à l'adresse

http://www.xlstat.com/demo-cochorcuttf.htm

Bibliographie

Cochrane D. and Orcutt G. (1949). Application of least squares regression to relationships containing autocorrelated error terms. *Journal of the American Statistical Association*, **44**, 32-61.

Tests d'hétéroscédasticité

Utilisez cet outil pour déterminer si des résidus obtenus à partir d'une régression linéaire peuvent être considérés comme ayant une variance indépendante de l'observation ou non.

Dans cette section :

Description

Le concept d' hétéroscédasticité - son contraire étant l'homoscédasticité - est utilisé en statistique, et plus particulièrement dans le contexte de la régression linéaire ou de l'étude de séries chronologiques, pour décrire le cas où la variance des erreurs du modèle n'est pas la même pour toutes les observations, alors que souvent, l''une des hypothèse de base en modélisation est que les variances sont homogènes et que les erreurs du modèle sont identiquement distribuées.

En régression linéaire, le fait que les erreurs (ou résidus) du modèle ne soient pas homoscédastiques a pour conséquence que les coefficients du modèle estimés par la méthode des moindres carrés ordinaires ne sont ni sans biais ni ceux de variance minimale et l'estimation de leur variance n'est pas fiable.

Il convient donc, si l'on soupçonne que les variances ne sont pas homogènes (une simple représentation des résidus en fonction des variables explicatives peut révéler une hétéroscédasticité), d'effectuer un test d'hétéroscédasticité. Plusieurs tests ont été mis au point, avec pour hypothèses nulle et alternative :

H0 : Les résidus sont homoscédastiques

Ha : Les résidus sont hétéroscédastiques

Test de Breusch-Pagan

Ce test a été mis au point par Breusch et Pagan (1979), puis amélioré par Koenker (1981) - le test est parfois appelé test de Breusch-Pagan et Koenker - pour permettre d'identifier des cas d'hétéroscédasticité rendant les estimateurs classiques des paramètres de la régression linéaire peu fiables. Si e désigne le vecteur des erreurs du modèle, l'hypothèse H0 peut s'écrire

$$H0 : Var(u/x) = \sigma^2$$

$$H0 : Var(u/x) = E(e^2/x) = E(e^2/x_1, x_2, ..., x_k) = E(e^2) = \sigma^2$$

Pour vérifier que les erreurs quadratiques sont indépendantes des variables explicatives, ce qui peut se traduire par de très nombreuses formes fonctionnelles, le plus simple et de faire une régression linéaire des erreurs quadratiques par les variables explicatives. Si les données sont homoscédastiques, le coefficient de détermination R^2 devrait alors ne pas être différent de 0. Si H0 est n'est pas rejetée on pourra conclure que l'hétéroscédasticité, si elle existe, ne prend pas la forme fonctionnelle retenue. La pratique montre que l'hétéroscédasticité n'est pas problématique si H0 n'est pas retenue. Si H0 est rejetée, il est vraisemblance qu'il y ait hétéroscédasticité et qu'elle prenne la forme fonctionnelle décrite ci-dessus.

La statistique utilisée pour le test, proposée par Koenker (1981) est :

$$LM = nR^2$$

LM est l'acronyme de *Lagrange multiplier*. Cette statistique présente l'avantage d'être asymptotiquement distribuée suivant une loi du χ^2 à p degrés de liberté, où p est le nombre de variables explicatives.

Si l'hypothèse nulle est rejetée, il conviendra de transformer les données avant de faire la régression, ou d'utiliser des méthodes de modélisation permettant de prendre en compte la variabilité de la variance.

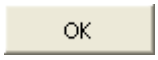
Test de White et test de White modifié (Wooldridge)

Ce test a été mis au point par White (1980) pour permettre d'identifier des cas d'hétéroscédasticité rendant les estimateurs classiques des paramètres de la régression linéaire peu fiables. L'idée est proche de celle de Breusch et Pagan, mais elle s'appuie sur des hypothèses plus faibles quand à la forme que prend l'hétéroscédasticité. Cela se traduit par une régression des erreurs quadratiques par les variables explicatives ainsi que par leurs carrés et leurs produits croisés (par exemple pour deux régresseurs, on prend x_1, x_2, x_1^2, x_2^2,x_1x_2 pour modéliser les erreurs quadratiques). La statistique utilisée est la même que pour le test de Breusch-Pagan, mais du fait de la présence de plus nombreux régresseurs, il y a cette fois 2p+p*(p-1)/2 degrés de liberté pour le Khi².

Afin d'éviter de perdre ces degrés de liberté, Wooldrigde (2009) propose de régresser les erreurs quadratiques par les prédictions du modèle et leur carré. On réduit ainsi à 2 le nombre de degrés de liberté pour le Khi².

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.

Onglet **Général**:

Résidus : Sélectionnez les résidus. Si le libellé de la colonne des résidus a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

X / Variables explicatives : sélectionnez la ou les variables quantitatives explicatives sur la feuille Excel. Les données sélectionnées doivent être de type numérique. Si le libellé des variables a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des variables : activez cette option si la première ligne des données sélectionnées (variables dépendantes et explicatives, libellés des observations, poids) contient un libellé.

Test de Breusch-Pagan : activez cette option pour effectuer un test de Breusch-Pagan.

Test de White : activez cette option pour effectuer un test de White. Activez l'option « Wooldridge » si vous voulez utiliser la version modifiée du test (voir la section description pour plus de détails).

Onglet **Options**:

Niveau de signification (%) : entrez la valeur du niveau de signification pour les tests (valeur par défaut : 5%).

Onglet **Données manquantes** :

Ne pas accepter les valeurs manquantes : activez cette option pour que XLSTAT empêche la poursuite des calculs si des valeurs manquantes sont détectées.

Supprimer les observations : activez cette option pour supprimer les observations comportant des données manquantes.

Onglet **Sorties**:

Statistiques descriptives : activez cette option pour afficher les statistiques descriptives des séries sélectionnées.

Onglet **Graphiques** :

Afficher les graphiques : activez cette option pour afficher le graphique présentant les résidus (en ordonnées) en fonction de la variable explicative (abscisse) si la sélection n'en comprend qu'une.

Résultats

Statistiques simples : dans ce tableau sont affichés, pour chacune des séries sélectionnées, le nombre d'observations, le nombre d'observations manquantes, le minimum, le maximum, la moyenne, et l'écart-type non biaisé.

Test de Breusch-Pagan : les résultats du test de Breusch-Pagan sont affichés si l'option correspondante a été activée, suivis d'une aide à l'interprétation.

Test de White : les résultats du test White sont affichés si l'option correspondante a été activée, suivis d'une aide à l'interprétation.

Exemple

Un exemple de test de racine unitaire est disponible sur le Centre d'aide XLSTAT :

http://www.xlstat.com/demo-whitetestf.htm

Bibliographie

Breusch T. and Pagan A. (1979). Simple test for heteroscedasticity and random coefficient variation. *Econometrica*, **47(5)**, 1287-1294.

Koenker R. (1981). A note on studentizing a test for heteroscedasticity. *Journal of Econometrics*, **17**, 107-112.

White H. (1980). A heteroskedasticity-consistant covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, **48(4)**, 817-838.

Wooldridge J.M. (2009). Introductory Econometrics. 4rth edition. Cengage Learning, KY, USA, 275-276.

Tests de racine unitaire et de stationnarité

Utilisez cet outil pour déterminer si une série temporelle est stationnaire ou non.

Dans cette section :

Description

Une série temporelle Y_t (t=1,2...) est dite stationnaire (au sens faible) si ses propriétés statistiques ne varient pas dans le temps (espérance, variance, auto-corrélation). Un exemple de série temporaire stationnaire est le bruit blanc, par exemple une série où la loi de Y_t est une loi normale $N(\mu, \sigma^2)$ indépendante de t. Un exemple de série non-stationnaire est la marche aléatoire définie par :

$$Y_t = Y_{t-1} + \epsilon_t$$

où ϵ_t est un bruit blanc

Identifier qu'une série n'est pas stationnaire permet ensuite d'étudier de quel type de non-stationnarité il s'agit. Une série non-stationnaire peut, entre autres, être **stationnaire en différence** : Y_t n'est pas stationnaire, mais différence mais $Y_t - Y_{t-1}$ est stationnaire. C'est le cas de la marche aléatoire. Une série peut également être **stationnaire en tendance**. C'est le cas de la série définie par

$Y_t = 0.5X_{t-1} + 1.4t + \epsilon_t$, où ϵ_t est un bruit blanc, qui n'est pas stationnaire.

En revanche, la série $Y_t - 1.4t = 0.5Y_{t-1} + \epsilon_t$ est stationnaire. Y_t est aussi stationnaire en différence.

Les tests de stationnarité permettent de vérifier si une série est stationnaire ou non. Il y a deux approches différentes : les tests pour lesquels l'hypothèse nulle H0 est que la série est stationnaire (test KPSS de Leybourne et McCabe), et ceux pour lesquels l'hypothèse nulle est au contraire que la série n'est pas stationnaire (test de Dickey-Fuller, test augmenté de Dickey-Fuller, test de Phillips-Perron, test DF-GLS). XLSTAT propose à ce jour le test KPSS, les tests de Dickey-Fuller et celui de Phillips-Perron.

Test de Dickey-Fuller

Ce test a été mis au point par Dickey et Fuller (1979) pour permettre d'identifier une racine unitaire dans une série temporelle pour laquelle on pense avoir qu'il y a un terme autorégressif d'ordre 1 et éventuellement un terme de tendance lié au temps. Pour rappel, le modèle autorégressif d'ordre 1 (noté AR(1)), peut s'écrire

$$X_t = \rho X_{t-1} + \epsilon_t \ , t = 1, 2, ...$$

où ϵ_t est une suite de variables indépendantes et identiquement distribuées suivant une loi normale $N(\mu, \sigma^2)$.

La série est stationnaire si $|\rho| < 1$. Elle n'est pas stationnaire et correspond à une marche aléatoire si $|\rho| = 1$.

Si l'on ajoute une constante et une tendance liée au temps, le modèle s'écrit

$$X_t = \rho X_{t-1} + \alpha + \beta t + \epsilon_t \ , t = 1, 2, ...$$

où ϵ_t est une suite de variables indépendantes et identiquement distribuées suivant une loi normale $N(\mu, \sigma^2)$.

Dickey et Fuller ont choisi de prendre comme hypothèse nulle $\rho = 1$ car elle a une implication opérationnelle immédiate : si l'hypothèse nulle n'est pas rejetée, alors pour analyser la série et éventuellement faire des prévisions, il sera nécessaire dans un premier temps de la transformer par différentiation (voir l'outil Transformation de séries temporelles, ou ARIMA).

Les deux hypothèses alternatives proposées, sont :

Ha(1) : $|\rho| < 1$, la série est stationnaire

Ha(2) : $|\rho| > 1$, la série est explosive

Les statistiques utilisées dans le test de Dickey-Fuller sont calculées à partir d'un modèle de régression linéaire et correspondent aux statistiques t du rapport entre un coefficient et son écart-type. Dickey et Fuller définissent :

- Modèle AR(1) :

$$\hat{\tau} = (\hat{\rho} - 1) / \sqrt{S_1^2 c_1}$$

- Modèle AR(1) avec la constante μ :

$$\hat{\tau}_\mu = (\hat{\rho}_\mu - 1) / \sqrt{S_2^2 c_2}$$

- Modèle AR(1) avec la constante μ et une tendance linéaire fonction de t :

$$\hat{\tau}_\tau = (\hat{\rho}_\tau - 1) / \sqrt{S_3^2 c_3}$$

Les S_k^2 correspondent à la moyenne des carrés des erreurs et les c_k à des variances.

Si ces statistiques sont simples à calculer, leurs distributions exacte et asymptotique sont complexes.

Les valeurs critiques ont été estimées au travers de simulations Monte Carlo par les auteurs avec plusieurs améliorations proposées au fil du temps.

MacKinnon (1996) a, quant à lui, mis au point des fonctions paramétriques approximant les distributions pour différentes tailles d'échantillon sur la base de très nombreuses simulations Monte Carlo. Pour l'estimation de la valeur critique et de la p-value, XLSTAT propose le choix entre une estimation à partir de simulations Monte Carlo et l'implémentation proposée par MacKinnon (1996). Dickey et Fuller, montrent que ces distributions ne dépendent pas de la distribution des ϵ_t et de la valeur initiale de la série Y_0.

Fuller (1976) avait déjà montré que cette approche peut-être généralisée à un modèle AR(p) pour déterminer s'il existe une racine unitaire sans pour autant identifier si la non-stationnarité provient d'un terme en particulier.

Test augmenté de Dickey-Fuller

Ce test a été mis au point par Said et Dickey (1984) et vient compléter le test de Dickey et Fuller en permettant de généraliser l'approche valable pour les modèles AR(p) à un modèle ARMA(p, q) pour lequel on va donc supposer qu'il est ARIMA(p, d, q), avec d=1 sous l'hypothèse H0. Said et Dickey montrent que l'on n'a pas besoin de connaître p, d et q pour appliquer le test de Dickey- Fuller présenté ci-dessus. Néanmoins, un paramètre k, correspondant à l'horizon de prise en compte de la partie moyenne mobile du modèle doit être fourni pour pouvoir effectuer le test. Par défaut, XLSTAT prend pour k la valeur suivante:

$$k = ENT((n-1)^{1/3})$$

où $ENT()$ est la partie entière

Said et Dickey montrent la statistique t du test de Dickey-Fuller peut être utilisée. Sa distribution asymptotique est la même que celle du test de Dickey-Fuller.

Test de Phillips- Perron

Une approche complémentaire à celle de Said et Dickey (1984) pour généraliser le test de Dickey et Fuller à différents processus générateurs de données a été proposée par Phillips (1987a) et développée par la suite dans Perron (1988) et Phillips et Perron (1988).

Comme pour le test de Dickey-Fuller (DF), le test de Phillips-Perron considère 3 régressions possibles sélectionnable dans XLSTAT : sans constante ni tendance temporelle, avec constante et sans tendance temporelle et avec constante et tendance temporelle. Les expressions des 3 modèles possibles sont données ci-dessous :

$$X_t = \rho X_{t-1} + \epsilon_t$$
$$X_t = \rho X_{t-1} + \alpha + \epsilon_t$$
$$X_t = \rho X_{t-1} + \alpha + \beta \cdot (t - T/2) + \epsilon_t$$

Il est à noter que dans le cas du test de Phillips-Perron, le terme d'erreur ϵ_t est supposé présenter une moyenne nulle mais peut avoir des corrélations sérielles (MA(q)) et être distribué de manière hétéroscédastique.

Contrairement au test augmenté de Dickey et Fuller (ADF), le test de Phillips-Perron (PP) ignore la corrélation sérielle au moment de l'ajustement du modèle. En revanche, une correction non paramétrique sur l'expression de la statistique est appliquée pour corriger l'effet que la corrélation sérielle et l'hétéroscédasticité du processus générateur pourraient avoir sur les résidus de l'ajustement. La statistique notée Z_τ est donnée par :

$$Z_t = \frac{\hat{\sigma}}{\hat{\lambda}} t_\rho - \frac{1}{2}\left(\frac{\hat{\lambda^2} - \hat{\sigma^2}}{\hat{\lambda}^2}\right)\left(\frac{T \times SE(\hat{\rho})}{\hat{\sigma}^2}\right)$$

où $\hat{\lambda}^2$ et $\hat{\sigma}^2$ sont des estimateurs convergents des paramètres de variance :

$$\hat{\lambda}^2 = \lim_{x \to +\infty} T^{-1} \sum_t^T E\left[T^{-1}(\sum_{t=1}^r \epsilon_t)^2\right]$$

$$\hat{\sigma^2} = \lim_{x \to +\infty} T^{-1} \sum_t^T E[\epsilon_t^2]$$

et

$$t_\rho = \frac{\hat{\rho} - 1}{SE(\hat{\rho})}$$

L'estimateur $\hat{\lambda}^2$ s'appuie sur l'estimateur de la matrice de covariance proposé par Newey et West (1987). Il permet de garantir la robustesse du résultat face aux corrélations sérielles et à l'hétéroscédasticité. Son implémentation est proposée dans deux configurations dans XLSTAT :

- Courte (par défaut): le nombre de pas considérés pour le calcul de l'estimateur de Newey-West est donnée par :

$$k = ENT(4(\frac{T}{100})^{2/9})$$

- Longue : pour les traces présentant un ordre MA plus élevé, le nombre de pas considérés est donnée par :

$$k = ENT(12(\frac{T}{100})^{2/9})$$

où $ENT()$ désigne la partie entière

Les statistiques ainsi obtenues peuvent être testées sur les mêmes distributions que celles du test de Dickey-Fuller. Les valeurs critiques et la p-value retournées sont donc celles obtenues à partir des fonctions paramétriques proposées par MacKinnon (1996) ou bien par simulation Monte Carlo.

Les avantages de test de Phillips-Perron face au test d'ADF sont d'autoriser l'hétéroscédasticité dans le processus générateur de ϵ_t et de ne pas nécessiter un paramétrage sensible de l'estimateur de Newey-West comme c'est le cas pour ADF.

Test KPSS de stationnarité

Ce test tient son nom de ses auteurs (Kwiatkowski, Phillips, Schmidt et Shin, 1991). Au contraire des tests de Dickey-Fuller, ce test permet de tester l'hypothèse nulle que la série est stationnaire. Soit le modèle

$$Y_t = \xi t + r_t + \epsilon_t, \ , t = 1, 2, ...$$

où ϵ_t est une erreur stationnaire, et r_t est une marche aléatoire définie par

$$r_t = r_{t-1} + u_t$$

où r_0 est une constante et où les u_t sont indépendantes et identiquement distribuées de moyenne 0 et de variance σ^2.

La série Y_t sera stationnaire dans le cas où la variance σ^2 est nulle. Elle sera alors stationnaire en tendance si ξ n'est pas nulle et en niveau (autour de r_0) si $\xi = 0$.

Soit n le nombre de pas de temps dont on dispose pour la série. Soit e_t les résidus du modèle de régression linéaire des Y_t sur le temps et une constante lorsque l'on veut tester la stationnarité en tendance, et les écarts à la moyenne pour la stationnarité en niveau.

On définit

$$s^2(l) = \frac{1}{n}\sum_{t=1}^{n} e_t^2 + \frac{2}{n}\sum_{s=1}^{l} w(s,l) \sum_{t=s+1}^{n} e_t e_{t-s}$$

avec

$$w(s,l) = 1 - s(l+1)$$

Soit S_t^2 la moyenne des carrés des erreurs entre 1 et t. La statistique pour le test de stationnarité en niveau est alors :

$$\eta_\mu = \frac{1}{n^2}\sum_{t=1}^{n} S_t^2 / s^2(l)$$

Pour le test de stationnarité en tendance on a :

$$\eta_\tau = \frac{1}{n^2}\sum_{t=1}^{n} S_t^2 / s^2(l)$$

la différence provenant des résidus.

Comme pour le test de Dickey-Fuller, ces statistiques sont simples à calculer mais leurs distributions exactes et asymptotiques sont complexes. Kwiatkowski *et al.* ont calculé les valeurs critiques asymptotiques sur la base de simulations Monte Carlo. XLSTAT permet de calculer les valeurs critiques et les p-values adaptées à la taille de l'échantillon, au travers de simulations Monte-Carlo réalisées à chaque utilisation.

Pondération par la méthode de Newey-West

L'estimateur de Newey-West (1987) est utilisé pour réduire l'effet de la dépendance (corrélation, autocorrélation) et de l'hétéroscédasticité (variances non homogènes) des erreurs d'un modèle. Le principe consiste à pondérer les erreurs du modèle pour le calcul des statistiques les impliquant. Si L est le nombre de pas pris en compte, le poids de chaque erreur est donné par :

$w_l = 1 - \frac{l}{L+1}$, l =1,2...,L

Le test KPSS implique des régressions qui présupposent l'homoscédasticité des erreurs. La pondération de Newey-West est proposée par XLSTAT pour si possible corriger une éventuelle hétéroscédasticité. XLSTAT recommande pour la détermination de L, deux possibilités :

- Court : $L = ENT(3\sqrt{n}/13)$

- Long : $L = ENT(10\sqrt{n}/14)$

où $ENT()$ est la partie entière.

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.

OK : cliquez sur ce bouton pour lancer les calculs.

Annuler : cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

Aide : cliquez sur ce bouton pour afficher l'aide.

 : cliquez sur ce bouton pour rétablir les options par défaut.

 : cliquez sur ce bouton pour effacer les sélections de données.

 : cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.

Onglet **Général**:

Séries temporelles : sélectionnez la ou les séries temporelles. Si des en- têtes de colonnes ont été sélectionnés, veuillez vérifier que l'option « Libellés des séries » est activée.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des séries : activez cette option si la première ligne des données sélectionnées (variables dépendantes et explicatives, libellés des observations, poids) contient un libellé.

Test de Dicker-Fuller : activez cette option pour effectuer un test de Dickey-Fuller. Choisissez ensuite le type de test à effectuer (voir la section description pour plus de détails).

Test de Phillips-Perron : activez cette option pour effectuer un test de Dickey-Fuller. Choisissez ensuite le type de test à effectuer (voir la section description pour plus de détails).

Test KPSS : activez cette option pour effectuer un test de KPSS. Choisissez ensuite le type de test à effectuer (voir la section description pour plus de détails).

Onglet **Options**:

Niveau de signification (%) : entrez la valeur du niveau de signification pour les tests (valeur par défaut : 5%).

Méthode : choisissez la méthode à utiliser pour le calcul de la p-value et de la valeur critique.

- Régression surfacique : cette option correspond à la méthode de calcul de MacKinnon (1996).
- Monte Carlo : cette option correspond à un calcul basé sur des simulations Monte Carlo.

Test de Dickey-Fuller : dans le cas d'un test de Dickey-Fuller augmenté, vous pouvez utiliser la valeur par défaut de k (voir la section « Description » pour plus de détails) ou entrer votre propre valeur.

Test de Phillips-Perron : dans le cas d'un test de Phillips-Perron, vous pouvez utiliser la valeur par défaut du nombre de pas (voir la section « Description » pour plus de détails) ou bien choisir la valeur longue.

Test de KPSS : choisissez si vous voulez utiliser la méthode de **Newey- West** pour supprimer l'impact de possibles autocorrélations sur les résidus du modèle. Pour le décalage utilisé, vous pouvez choisir entre **court**, **long**, ou vous pouvez entrer votre propre valeur pour L (voir la section « Description » pour plus de détails).

Onglet **Données manquantes** :

Ne pas accepter les valeurs manquantes : activez cette option pour que XLSTAT empêche la poursuite des calculs si des valeurs manquantes sont détectées.

Supprimer les observations : activez cette option pour supprimer les observations comportant des données manquantes.

Remplacer par la moyenne des valeurs précédente et suivante : activez cette option pour estimer les données manquantes par la moyenne de la première valeur précédente non manquante et de la première valeur suivante non manquante.

Ignorer les données manquantes : activez cette option pour ignorer les données manquantes.

Onglet **Sorties**:

Statistiques descriptives : activez cette option pour afficher les statistiques descriptives des séries sélectionnées.

Résultats

Statistiques simples : dans ce tableau sont affichés, pour chacune des séries sélectionnées, le nombre d'observations, le nombre d'observations manquantes, le minimum, le maximum, la moyenne, et l'écart-type non biaisé.

Test de Dicker-Fuller : les résultats du test de Dickey-Fuller sont affichés si l'option correspondante a été activée, suivis d'une aide à l'interprétation.

Test de Phillips -Perron : les résultats du test de Phillips-Perron sont affichés si l'option correspondante a été activée, suivis d'une aide à l'interprétation.

Test KPSS : les résultats du test KPSS sont affichés si l'option correspondante a été activée, suivis d'une aide à l'interprétation.

Exemple

Un exemple de test de racine unitaire est disponible sur le Centre d'aide XLSTAT :

http://www.xlstat.com/demo-unitrootf.htm

Bibliographie

Brockwell P.J. and Davis R.A. (1996). Introduction to Time Series and Forecasting. Springer Verlag, New York.

Dickey D. A. and Fuller W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association*, **74(366)**, 427-431.

Fuller W.A. (1996). Introduction to Statistical Time Series, Second Edition. John Wiley & Sons, New York.

Kwiatkowski D. , Phillips P. C. B., Schmidt P. and Y. Shin (1992). Testing the null hypothesis of stationarity against the alternative of a unit root. *Journal of Econometrics,* **54**, 159-178.

MacKinnon J. G. (1996). Numerical distribution functions for unit root and cointegration tests. *Journal of Applied Econometrics*, **11**, 601-18.

Newey W. K. and West K. D (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, **55(3)**, 703-708.

Said S. E. and Dickey D. A. (1984). Testing for Unit Roots in Autoregressive-Moving Average Models of Unknown Order. *Biometrika*, **71**, 599-607.

Phillips P. C. B. (198 7). Time series regression with a unit root. *Journal of the Economic Society*, 277-301.

Perron P. (198 8). Trends and random walks in macroeconomic time series: Further evidence for a new approach. *Journal of economic dynamics and control*, **12(2)**, 297-332.

Phillips P. C. B. and Perron P. (198 8). Testing for a unit root in time series regression. *Biometrika*, **75(2)**, 335-346.

Tests de cointégration

Utilisez cet outil pour réaliser un test de cointégration selon l'approche de Johansen (1991, 1995) sur un groupe de deux ou plus séries temporelles.

Dans cette section :

Description

La théorie économique suggère souvent des relations de long terme entre 2, ou plus, variables économiques. Ainsi, bien que ces variables puissent dévier les unes des autres pendant de courtes périodes de temps, les forces économiques à l'œuvre vont avoir tendance, sur le long terme, à restaurer l'équilibre qui existait à l'origine. Des exemples de telles relations peuvent se retrouver en économie entre la monnaie, les revenus, les prix et les taux d'intérêts à long terme ou bien encore entre les taux d'échange, les prix extérieurs et les prix domestiques. En finance, de telles relations sont attendues, par exemple, entre les prix d'un même actif sur différentes places boursières.

Le terme de cointégration a été introduit pour la première fois dans Engle and Granger (1987) après le travail publié dans Granger and Newbold (1974) sur les régressions erronées. Il identifie ce type de situation où 2, voire d'avantage, séries chronologiques non stationnaires sont liées de telle manière qu'elles ne peuvent pas dévier les unes des autres sur le long terme. Il existe alors une ou plusieurs combinaisons linéaires de ces séries temporelles intégrées d'ordre 1 (ou I(1), voir Test de racine unitaire) qui soient stationnaires (ou I(0)). Ces combinaisons linéaires stationnaires sont nommées équations de cointégration.

L'une des approches les plus intéressantes pour tester la cointégration d'un groupe de séries temporelles est la méthode du maximum de vraisemblance proposée par Johansen (1988, 1991). Cette approche, implémentée dans XLSTAT, est basée sur le modèle Vecteur Autorégressif (VAR) et peut être décrite comme suit.

Tout d'abord, considérons la représentation en niveau d'un modèle $VAR(P)$, pour Y_t, un K-vecteur de séries temporelles I(1) :

$$Y_t = \Phi D_t + \Pi_1 Y_{t-1} + ... + \Pi_P Y_{t-P} + \epsilon_t \text{ pour } t = 1, ..., T$$

Avec D_t le vecteur contenant les composantes tendancielles comme une constante ou une tendance temporelle et ϵ_t le vecteur d'innovations.

Le paramètre P définit l'ordre du modèle VAR et est l'un des paramètres d'entrée de la méthode de Johansen relative aux tests de cointégration. Si vous ne connaissez pas la valeur que devrait prendre ce paramètre, sélectionnez l'option automatique dans l'onglet Général. Vous pourrez alors spécifier dans l'onglet Options le modèle de VAR qui décrit le mieux vos données (ni constante ni tendance, constante, tendance, constante+tendance), choisir le critère de sélection parmi les 4 proposés (AIC, FPE,HQ et BIC) et saisir une valeur maximale pour le nombre de décalages évalués. XLSTAT estimera alors la valeur du paramètre P selon la procédure détaillée dans LüktePohl (2005) et réalisera ensuite le reste de l'analyse. Le détail des résultats pour l'estimation du modèle VAR est affiché en fin d'analyse pour permettre une vérification par l'utilisateur.

Selon le théorème de représentation de Granger, un modèle $VAR(P)$ avec des variables I(1) peut être représenté de manière équivalente par un modèle à correction d'erreur (VECM) :

$$\Delta Y_t = \Phi D_t + \Pi Y_{t-1} + \Gamma_1 \Delta Y_{t-1} + ... + \Gamma_{P-1} \Delta Y_{t-P+1} + \epsilon_t$$

Avec Δ l'opérateur de différenciation, $\Pi = \Pi_1 + ... + \Pi_{P-1} - I_K$ et $\Gamma_l = -\sum_{j=l+1}^{P} \Pi_j$ pour $l = 1, ..., P-1$

Dans cette représentation, le terme ΔY_t et ses décalages sont tous I(0). Le terme Y_{t-1} est donc la seule composante potentiellement non-stationnaire. Ainsi, pour que l'équation ci-dessus soit valide (une combinaison linéaire de composantes I(0) est elle-même I(0)), le terme ΠY_{t-1} doit contenir les relations de cointégration si elles existent.

Trois cas peuvent être considérés :

- La matrice Π est égale à 0 ($rang(\Pi) = 0$), il n'y a alors pas de relation de cointégration,
- La matrice Π est de rang plein ($rang(\Pi) = K$), toutes les composantes de Y_{t-1} sont indépendamment stationnaires (ce qui est en désaccord avec l'hypothèse de départ de séries I(1)),
- La matrice Π n'est ni nulle ni de rang plein ($0 < rang(\Pi) < K$), Y_{t-1} est alors I(1) avec r vecteurs cointégrants indépendants et $K - r$ tendances stochastiques communes.

Dans ce dernier cas, la matrice Π peut s'écrire comme le produit suivant :

$$\Pi_{(K \times K)} = \alpha_{(K \times r)} \beta_{(r \times K)}$$

Avec $rang(\alpha) = rang(\beta) = r$.

La matrice β est la matrice cointégrante et ses colonnes forment une base pour les coefficients de cointégration. La matrice α, souvent appelée matrice d'impact ou de chargement, contrôle la

rapidité avec laquelle les effets induits par Y_{t-1} se propagent à ΔY_t. Il est important de noter que la factorisation $\Pi = \alpha\beta'$ n'est pas définie de manière unique. Une normalisation supplémentaire peut être nécessaire pour obtenir des estimations uniques de α et β. Les valeurs reportées dans XLSTAT utilisent la normalisation proposée par Johansen (1995) :$\beta' \cdot S_{11} \cdot \beta = I_r$

La méthodologie de test estime la matrice Π et contruit ensuite des tests de rapport de vraisemblance successifs pour estimer son rang à partir de ses valeurs propres $\widehat{\lambda_1} > \widehat{\lambda_2} > \ldots > \widehat{\lambda_x}$. Le rang de Π est égal au nombre de valeurs propres significativement différentes de 0. Il donne également le rang de cointégration du système (ou, de manière équivalente, le nombre d'équations cointégrantes).

Deux procédures séquentielles sont proposées par Johansen pour évaluer le rang de cointégration r_0:

- Le test de λ_{max}(ou lambda max) utilise la statistique suivante : $LR_{max}(r_0) = -T\ln(1-\hat{\lambda}_{r_0+1})$,
- Le test de trace pour lequel la statistique est $LR_{trace}(r_0) = -T\sum_{i=r_0+1}^{n}\ln(1-\hat{\lambda}_i)$.

En commençant par l'hypothèse nulle selon laquelle il n'existe pas de relation de cointégration ($r_0 = 0$, le test de lambda max évalue si la $(r_0+1)^{ème}$ valeur propre peut être considérée comme nulle. Si l'hypothèse $\lambda_{r_0+1} \approx 0$ est rejetée, on procède au test du prochain niveau de cointégration. De manière similaire, LR_{trace} pour le test de trace est proche de zéro si le $rang(\Pi) = r_0$ et large si $rang(\Pi) > r_0$.

Les distributions asymptotiques de ces tests de vraisemblance sont non-standards et dépendent des hypothèses prises sur les composantes déterministes de ΔY_t :

$$\Delta Y_t = c_1 + d_1 t + \alpha(\beta' Y_{t-1} + c_0 + d_0 t) + \Gamma_1 \Delta Y_{t-1} + ... + \Gamma_{P-1}\Delta Y_{t-P+1} + \epsilon_t$$

Cinq types de restriction sont considérés selon les tendances identifiées pour Y_t et ΔY_t :

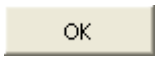
- **H2** $(c_0 = c_1 = d_0 = d_1 = 0)$: les séries de Y_t sont I(1) sans tendance déterministe en niveau et celles de $\beta' Y_t$ possèdent une moyenne nulle. Dans la pratique, ce cas est rarement rencontré.
- **H1*** $(c_1 = d_0 = d_1 = 0)$: les séries de Y_t sont I(1) sans tendance déterministe en niveau et celles de $\beta' Y_t$ possèdent une moyenne non nulle.
- **H1** $(d_0 = d_1 = 0)$: les séries de Y_t sont I(1) avec une tendance linéaire en niveau et celles de $\beta' Y_t$ possèdent une moyenne non nulle.
- **H*** $(d_1 = 0)$: les séries de Y_t et $\beta' Y_t$ possèdent une tendance linéaire.
- **H** (non-contrainte): les séries de Y_t sont I(1) avec une tendance quadratique en niveau et celles de $\beta' Y_t$ ont une tendance linéaire. Ce cas est, également, rarement rencontré en pratique.

Pour effectuer un test de cointégration dans XLSTAT, il faut sélectionner l'une des contraintes ci-dessus. Ce choix devrait être motivé par des considérations sur la nature spécifique de votre jeu de données ou bien sur le modèle économique à l'œuvre. Cependant, si le choix d'une restriction n'est pas clair pour votre problème, une bonne approche consiste à évaluer la robustesse des résultats en sélectionnant successivement les hypothèses $H1*$, $H1$ et $H*$ (les 2 options restantes étant très spécifiques et facilement identifiables).

Les résultats des tests de lambda max et de trace (valeurs critiques et p-values) affichés par XLSTAT sont estimés à partir des travaux publiés dans MacKinnon-Haug-Mechelis (1998).

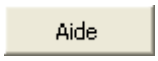
Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.



: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.

Onglet **Général** :

Séries temporelles : sélectionnez la ou les séries temporelles. Si des en- têtes de colonnes ont été sélectionnés, veuillez vérifier que l'option « Libellés des séries » est activée.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des séries : activez cette option si la première ligne des données sélectionnées (variables dépendantes et explicatives, libellés des observations, poids) contient un libellé.

Modèle : sélectionnez le type de restriction qui s'applique le mieux à vos données entre $H2$, $H1*$, $H1$, $H*$ et H (voir la description pour plus de détails).

Ordre du VAR : sélectionnez automatique pour qu'XLSTAT estime le paramètre P (voir la description pour plus de détails) ou bien sélectionnez défini par l'utilisateur et saisissez votre propre valeur de paramètre.

Onglet **Options** :

Niveau de signification (%) : entrez la valeur du niveau de signification pour les tests (valeur par défaut : 5%).

Estimation de l'ordre du VAR : si l'option automatique est sélectionnée dans l'onglet Général, vous devez choisir 3 paramètres pour estimer l'ordre du VAR : le modèle, le critère de sélection et le nombre maximum de décalage.

Modèle : sélectionnez entre.aucun, constante, tendance et constante+tendance le modèle qui s'applique le mieux à vos données.

Critère de sélection : sélectionnez le critère parmi les 4 calculés (AIC, FPE, HQ, BIC) qui sera utilisé par XLSTAT pour choisir l'ordre du VAR.

Nombre maximum de décalage : sélectionnez le nombre maximum de décalage à évaluer.

Onglet **Données manquantes** :

Ne pas accepter les valeurs manquantes : activez cette option pour que XLSTAT empêche la poursuite des calculs si des valeurs manquantes sont détectées.

Supprimer les observations : activez cette option pour supprimer les observations comportant des données manquantes.

Remplacer par la moyenne des valeurs précédente et suivante : activez cette option pour estimer les données manquantes par la moyenne de la première valeur précédente non manquante et de la première valeur suivante non manquante.

Onglet **Sorties** :

Statistiques descriptives : activez cette option pour afficher les statistiques descriptives des séries sélectionnées.

Résultats

Statistiques simples : dans ce tableau sont affichés, pour chacune des séries sélectionnées, le nombre d'observations, le nombre d'observations manquantes, le minimum, le maximum, la moyenne, et l'écart-type non biaisé.

Estimation de l'ordre du VAR : si l'option automatique est sélectionnée pour l'estimation de l'ordre du VAR. Chaque ligne correspond à l'évaluation des 4 critères de sélection pour un nombre de décalage allant de 1 au nombre maximum de décalage. Le critère discriminant sélectionné est en gras.

Test de lambda max : ce tableau affiche, pour chaque rang de cointégration testé, la valeur propre correspondante, la statistique du test de lambda max et les valeurs critiques et p-values.

Test de la trace : ce tableau affiche, pour chaque rang de cointégration testé, la valeur propre correspondante, la statistique du test de trace et les valeurs critiques et p-values.

Coefficients d'ajustement (alpha) : ce tableau affiche la matrice d'impact ou matrice de chargement (voir la description pour plus de détails).

Coefficients de cointégration (beta) :ce tableau affiche la matrice cointégrante (voir la description pour plus de détails).

Exemple

Un exemple de test de cointégration est disponible sur le Centre d'aide XLSTAT :

http://www.xlstat.com/demo-cointegrationf.htm

Bibliographie

Engle R.and Granger C. (1987). Co-integration and error correction: Representation, estimation and testing. Econometrica: Journal of the Econometric Society, pp.251-276.

Granger C. and Newbold P. (1974). Spurious regressions in econometrics. Journal of econometrics, 2(2), pp.111-120.

Johansen, S. (1988). Statistical analysis of cointegration vectors. Journal of economic dynamics and control, 12(2), pp.231-254.

Johansen S. (1991). Estimation and Hypothesis Testing of Cointegration Vectors in Gaussian Vector Autoregressive Models. Econometrica: Journal of the Econometric Society, pp.1551-1580.

Johansen S. (1995). Likelihood based inference in cointegrated vector autoregressive models. OUP catalogue.

Lüktepohl (2005). New introduction to multiple time series analysis. Springer.

MacKinnon, J. G., Haug, A. A., & Michelis, L. (1998). Numerical distribution functions of likelihood ratio tests for cointegration(No. 9803). Department of Economics, University of Canterbury.

Transformation de séries temporelles

Utilisez cet outil pour transformer une série en une nouvelle série ayant de meilleures propriétés : tendance et saisonnalité retirées, normalité et stationnarité accrues.

Dans cette section :

Description

XLSTAT offre plusieurs possibilités pour transformer une série X_t en une série Y_t, $(t = 1, \ldots, n)$:

Transformation Box-Cox : elle permet d'augmenter la normalité des données. L'équation de Box-Cox est définie par :

$$Y_t = \begin{cases} \frac{X_t^\lambda - 1}{\lambda}, & (X_t > 0,\ \lambda \neq 0)\ ou\ (X_t \geq 0,\ \lambda > 0) \\ ln(X_t), & (X_t > 0, \lambda = 0) \end{cases}$$

XLSTAT accepte soit une valeur fixée pour λ, soit de trouver la valeur optimale permettant de maximiser la vraisemblance pour le modèle linéaire simple ayant le temps pour variable explicative.

Différenciation : permet de supprimer les tendances et la saisonnalité, et d'obtenir la stationnarité des séries. L'équation de différenciation est donnée par :

$$Y_t = (1 - B)^d (1 - B^s)^D X_t$$

où d est l'ordre de différenciation pour la composante tendancielle, s est la période de la composante saisonnière, et D est l'ordre de la composante saisonnière. B est l'opérateur mathématique de décalage, défini par :

$$BX_t = X_{t-1}$$

Les valeurs de (d, D, s) peuvent être choisies par essais successifs, ou suggérées par l'analyse descriptive des séries (fonctions FAC ou FACP par exemple). Des valeurs communes sont $(1,0,0)$, $(1,1,s)$, $(2,1,s)$. s vaut 12 pour des données mensuelles avec une saisonnalité annuelle, 0 lorsqu'il n'y a pas de saisonnalité.

Detrending et **désaisonnalisation** par utilisation du modèle classique de décomposition donné par :

$$X_t = m_t + s_t + \epsilon_t$$

où m_t est la composante tendancielle, et s_t la composante saisonnière et ϵ_t un bruit blanc suivant une loi $N(0,1)$. XLSTAT permet d'ajuster ce modèle en deux étapes séparées ou successives :

1 – Ajustement du modèle de detrending suivant :

$$X_t = m_t + \epsilon_t = \sum_{i=0}^{k} a_i t^i + \epsilon_t$$

où k est le degré du polynôme. Les paramètres a_i sont obtenus par ajustement d'un modèle linéaire sur les données. La série transformée s'écrit :

$$Y_t = \epsilon_t = X_t - \sum_{i=0}^{p} a_i t^i$$

2 – Ajustement d'un modèle de désaisonnalisation :

$$X_t = s_t + \epsilon_t = \mu + b_i + \epsilon_t, \quad i = t \bmod p$$

où p est la période. Les paramètres b_i sont obtenus par ajustement d'un modèle linéaire aux données. La série transformée est donnée par :

$$Y_t = \epsilon_t = X_t - b_i - \mu$$

Remarque : il existe de nombreuses autres transformations possibles. Des filtres linéaires peuvent aussi être utilisés. Un lissage par moyenne mobile peut être utilisé pour filtrer des bruits. Les méthodes de lissage sont proposées dans la section Lissage.

Décomposition saisonnière permet de calculer les indices saisonniers et de décomposer la série temporelle en 3 composantes (tendancielle, saisonnière et aléatoire) à partir de la période P définie par l'utilisateur.

Dans le cas d'une décomposition choisie de type additif, le modèle s'exprime comme suit :

$$X_t = m_t + s_{t \bmod p} + \epsilon_t$$

Où X_t est la série temporelle considérée, m_t est la composante tendancielle, $s_{t \bmod p}$ la composante saisonnière et ϵ_t la composante aléatoire.

La composante tendancielle est estimée au moyen d'un filtre à moyenne mobile centrée :

$$\hat{m_t} = \sum_{i=-P/2}^{P/2} W_i \times X_{t+i}$$

Avec $P/2$ la division entière de P par 2 et les coefficients w_i définis par :

$$W_i = \begin{cases} \frac{1}{2P}, & si \mid i \mid = P/2 \\ \\ \frac{1}{P}, & autrement \end{cases}$$

Chaque indice saisonnier s_i est ensuite estimé à partir de la différence $s_t = X_t - m_t$ comme étant la moyenne arithmétique des éléments de s_t pour lesquels $t \bmod P = i$.

La valeur de ces indices est ensuite recentrée comme indiqué ci-dessous :

$$\hat{s}_i = \hat{s}_i - \frac{1}{P} \sum_{j=1}^{p} \hat{s}_j$$

Enfin, la composante aléatoire est estimée par l'équation suivante :

$$\hat{\epsilon}_i = X_t - \hat{m}_t - \hat{s}_{t \bmod P}$$

Dans le cas d'un modèle de type multiplicatif, le modèle s'exprime comme suit :

$$X_t = m_t \times s_{t \bmod P} \times \epsilon_t$$

La composante tendancielle est estimée de la même manière que pour le modèle additif.

En revanche, les indices saisonniers sont estimés par la moyenne arithmétique des éléments du rapport $s_t = X_t / m_t$ pour lesquels $t \bmod P = i$

Ils sont ensuite normalisés comme suit :

$$\hat{S_i} = \hat{s_i} \times (\prod_{j=1}^{p} \hat{s}_j)^{-1/P}$$

Enfin, la composante aléatoire est estimée par l'équation suivante :

$$\hat{\epsilon_i} = \frac{X_t}{\hat{s}_{t \bmod P} \times \hat{m}_t}$$

La **décomposition en modes empiriques** (EMD) permet de décomposer tout signal aussi complexe soit-il en un nombre fini et souvent faible de composantes appelées IMFs, pour les fonctions de mode intrinsèque. La transformation de Hilbert de ces composantes permet d'obtenir leurs fréquences instantanées qui permettent des interprétations physiques pertinentes.

Les avantages de cette méthode sont les suivants : * Elle peut être utilisée sur des séries non stationnaires et non linéaires * Elle est adaptative, puisque les composantes sont déduites des données elles mêmes, ce qui la rend très efficace

Le coeur de cet algorithme, qui permet d'extraire les IMFs, est la procédure dite de "tamisage", ou sifting en anglais. Elle peut être décrite comme suit : * Identifier les extrema locaux the signal d'entrée S * Calculer les enveloppes supérieures et inférieures du signal par interpolation par spline cubique, puis en calculer la moyenne $$ m(t) = \frac{e_{upper}(t) + e_{lower}(t)}{2} $$ * Soustraire la moyenne des deux enveloppes au signal $$ h(t) = S(t) - m(t) $$ * Répéter ces étapes sur $h(t)$ jusqu'à ce qu'un critère d'arrêt soit atteint. Le signal obtenu $\hat{h}_1$ peut être considéré comme étant une IMF.

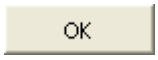
Pour trouver les prochaines IMFs, répéter la procédure sur le résidu, défini comme suit : $$ S_{i+1}(t) = S_i(t) - \hat{h}_{i}(t) $$

La **décomposition en modes empiriques d'ensemble** (EEMD) a été inventée en 2009 comme une méthode d'analyse de données assistée par bruit blanc, et dont le but est de résoudre le problème de mélange de modes qui pouvait être rencontré lors de l'utilisation de l'EMD classique. En effet, lors une simple EMD sur un signal contenant des composantes intermittentes, une IMF peut parfois contenir de l'information concernant des fréquences différentes.

Le principe est le suivant : * Ajouter un bruit blanc gaussien au signal * Extraire les IMFs de ce nouveau signal en utilisant une EMD classique * Répéter ces étapes N fois * Agréger les IMFs par la moyenne d'ensemble

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.

Onglet **Général**:

Séries temporelles : sélectionnez la ou les séries temporelles que vous voulez analyser. Si des en-têtes de colonnes ont été sélectionnés, veuillez vérifier que l'option « Libellés des séries » est activée.

Données de date : activez cette option pour sélectionner des données de date. Ces données doivent être au format de data Excel, ou des valeurs numériques.

Vérifier les intervalles : activez cette option si vous voulez que XLSTAT vérifie que les données de date sont bien régulièrement espacées.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des séries : activez cette option si la première ligne des données sélectionnées (variables dépendantes et explicatives, libellés des observations, poids) contient un libellé.

Onglet **Options**:

Transformation Box-Cox : activez cette option pour faire une transformation de Box-Cox. Vous pouvez soit imposer une valeur de **Lambda**, soit décider que XLSTAT doit l'**optimiser** (voir la description pour plus de détails).

Différenciation : activez cette option pour calculer la série différenciée. Vous devez saisir les valeurs des paramètres (**d, D, s**). Voir la description pour plus de détails.

Régression polynomiale : activez cette option pour retirer la composante tendancielle d'une série chronologique. Vous devez saisir le **degré du polynôme**. Voir la description pour plus de détails.

Ajustement saisonnier : activez cette option pour retirer la composante saisonnière d'une série chronologique. Vous devez saisir la **période**. Voir la description pour plus de détails.

Décomposition saisonnière : activez cette option pour calculer les indices saisonniers et décomposer la serie chronologique en entrée. Vous devez choisir le type de modèle, additif ou multiplicatif et saisir la **période**. Voir la description pour plus de détails.

Onglet **Données manquantes** :

Ne pas accepter les valeurs manquantes : activez cette option pour que XLSTAT empêche la poursuite des calculs si des valeurs manquantes sont détectées.

Supprimer les observations : activez cette option pour supprimer les observations comportant des données manquantes.

- **Vérifier pour chaque série séparément** : choisissez cette option si vous voulez que lorsque, pour une observation donnée, il y a des données manquantes uniquement dans les Y, l'observation ne soit supprimée que si la donnée correspondant au Y en cours de modélisation est manquante.
- **Pour tous les Y** : choisissez cette option pour supprimer toutes les observations pour lesquelles des Y sont manquants.
- Remarque : les deux alternatives ci-dessus sont sans effet si il n'y a qu'un seul Y.

Remplacer par la moyenne des valeurs précédente et suivante : activez cette option pour estimer les données manquantes par la moyenne de la première valeur précédente non manquante et de la première valeur suivante non manquante.

Onglet **Sorties**:

Statistiques descriptives : activez cette option pour afficher les statistiques descriptives des séries sélectionnées.

Paramètres du modèle : activez cette option pour afficher la valeur des paramètres du modèle après ajustement.

Coefficients d'ajustement : activez cette option pour afficher le tableau des statistiques d'ajustement du modèle.

Onglet **Graphiques** :

Afficher les graphiques : activez cette option pour afficher les graphiques permettant de comparer les séries avant et après transformation.

Résultats

Statistiques simples : dans ce tableau sont affichés, pour chacune des séries sélectionnées, le nombre d'observations, le nombre d'observations manquantes, le minimum, le maximum, la moyenne, et l'écart-type non biaisé.

Transformation de Box-Cox :

Paramètres du modèle : ce tableau n'est affiché que si l'option d'optimisation de Lambda a été choisie. Il présente les estimateurs des trois paramètres du modèle, qui sont Lambda, la constante du modèle linéaire, et le coefficient de pente.

Série avant et après transformation : dans ce tableau sont affichées la série avant transformation et la série après transformation. Si Lambda a été optimisé, la série après optimisation correspond aux résidus du modèle. Si Lambda est fixé, la série après transformation correspond à l'application directe de la transformation de Box-Cox.

Différenciation :

Série avant et après transformation : dans ce tableau sont affichées la série avant transformation et la série après transformation. Les $d + D + s$ premières observations ne sont pas affichées pour la série transformée en raison des contraintes liées à la méthode.

Régression polynomiale :

Coefficients d'ajustement : dans ce tableau sont affichés les coefficients d'ajustement du modèle polynomial.

Paramètres du modèle : dans ce tableau sont affichés les estimateurs des paramètres du modèle.

Série avant et après transformation : dans ce tableau sont affichées la série avant transformation et la série après transformation. La série après transformation correspond aux résidus du modèle.

Désaisonnalisation :

Coefficients d'ajustement : dans ce tableau sont affichés les coefficients d'ajustement du modèle polynomial.

Paramètres du modèle : dans ce tableau sont affichés les estimateurs des paramètres du modèle.

Série avant et après transformation : dans ce tableau sont affichées la série avant transformation et la série après transformation. La série après transformation correspond aux résidus du modèle.

Exemple

Un exemple de transformation de séries chronologiques est disponible sur le Centre d'aide XLSTAT à l'adresse suivante :

http://www.xlstat.com/demo-descf.htm

Bibliographie

Box G. E. P. and Jenkins G. M. (1976). Time Series Qnalysis: Forecasting and Control. Holden-Day, San Francisco.

Brockwell P.J. and Davis R.A. (1996). Introduction to Time Series and Forecasting. Springer Verlag, New York.

Shumway R.H. and Stoffer D.S. (2000). Time Series Analysis and Its Applications. Springer Verlag, New York.

N. E. Huang and al. (1998) “The empirical mode decomposition method and the Hilbert spectrum for non-stationary time series analysis”, Proc. Roy. Soc. London A.454

N. E. Huang and Z. Wu. (2009) “Ensemble empirical mode decomposition : a noise-assisted data analysis method”, Advanced Adaptive Data Analysis 1.1 (2009)

Lissage

Utilisez cet outil pour lisser une série temporelle et pour éventuellement prévoir des valeurs futures de la série.

Dans cette section :

Description

Plusieurs méthodes de lissage sont disponibles. On définit par Y_t, (t=1,…,n), la série temporelle étudiée, par $P_t\ Y_{t+h}$ la prévision de Y_{t+h} qui minimise la moyenne du carré des erreurs (MCE) et par ϵ_t un bruit blanc distribué suivant une loi normale $N(0,1)$. Les méthodes de lissage sont définies par les équations suivantes :

Lissage exponentiel simple

Ce modèle est aussi parfois appelé lissage exponentiel simple de Brown, ou le modèle à moyenne mobile exponentiellement pondérée. Les équations du modèle sont données par :

$$\begin{cases} Y_t = \mu_t + \epsilon_t & \\ P_t Y_{t+h} = \mu_t & h = 1, 2, ... \\ S_t Y_{t+h} = \alpha Y_t + (1-\alpha) S_{t-1} & 0 < \alpha < 2 \\ \hat{Y}_{t+h} = P_t Y_{t+h} = S_t, & h = 1, 2, ... \end{cases}$$

Le domaine de définition donné pour α correspond au domaine d'additivité et d'inversibilité du modèle.

Le lissage exponentiel simple permet de prédire une valeur en fonction des données passées, en donnant aux données un poids d'autant plus faible qu'elles correspondent à un passé éloigné. La pondération évolue de façon exponentielle, d'où le nom du modèle. En matière de prévision, ce modèle est assez limité, puisque les prévisions sont constantes au-delà de $n+1$.

Lissage exponentiel double

Ce modèle est parfois appelé Lissage exponentiel double de Brown ou lissage exponentiel linéaire de Brown. Les prévisions tiennent ici compte d'une tendance observée sur les données

précédentes. Les équations du modèle sont données par :

$$\begin{cases} Y_t = \mu_t + \beta_t t + \epsilon_t & \\ P_t Y_{t+h} = \mu_t + \beta_t t & h = 1, 2, ... \\ S_t = \alpha Y_t + (1-\alpha) S_{t-1} & 0 < \alpha < 2 \\ T_t = \alpha S_t + (1-\alpha) S_{t-1} & \\ \hat{Y}_{t+h} = P_t Y_{t+h} = (2 + \frac{\alpha h}{1-\alpha}) S_t - (1 + \frac{\alpha h}{1-\alpha}) T_t, & \alpha \neq 1, h = 1, 2, ... \\ \hat{Y}_{t+h} = P_t Y_{t+h} = Y_t, & \alpha = 0, h = 1, 2, ... \end{cases}$$

Le domaine de définition donné pour α correspond au domaine d'additivité et d'inversibilité du modèle.

Lissage exponentiel linéaire de Holt

Ce modèle est parfois appelé algorithme non-saisonnier de Holt-Winters. Comme le précédent, il permet de prendre en compte une composante tendancielle, mais avec plus de souplesse, car il fait intervenir un paramètre de plus. Les prévisions pour t>n prennent en compte la composante tendancielle. Les équations du modèle sont données par :

$$\begin{cases} Y_t = \mu_t + \beta_t t + \epsilon_t & \\ P_t Y_{t+h} = \mu_t + \beta_t t & h = 1, 2, ... \\ S_t = \alpha Y_t + (1-\alpha)(S_{t-1} + T_{t-1}) & 0 < \alpha < 2 \\ T_t = \beta (S_t - S_{t-1}) + (1-\beta) T_{t-1} & 0 < \beta < 4/\alpha - 2 \\ \hat{Y}_{t+h} = P_t Y_{t+h} = S_t + h T_t, & h = 1, 2, ... \end{cases}$$

Les domaines de définition donnés pour α et β correspondent au domaine d'additivité et d'inversibilité du modèle.

Modèle de Holt-Winters saisonnier additif

Cette méthode permet de prendre en compte une tendance qui varie avec le temps, et une composante saisonnière de période p. Les prévisions tiennent compte de la tendance et de la saisonnalité. Ce modèle met en jeu trois paramètres. On l'appelle additif car la composante saisonnière est stable dans le temps. Les équations du modèle sont données par :

$$\begin{cases} Y_t = \mu_t + \beta_t t + s_p(t) + \epsilon_t & \\ P_t Y_{t+h} = \mu_t + \beta_t t + s_p(t) & h = 1, 2, ... \\ S_t = \alpha (Y_t - S_{t-p} + (1-\alpha)(S_{t-1} + T_{t-1}) & \\ T_t = \beta (S_t - S_{t-1}) + (1-\beta) T_{t-1} & \\ D_t = \gamma (Y_t - S_t) + (1-\gamma) D_{t-p} & \\ \hat{Y}_{t+h} = P_t Y_{t+h} = S_t + h T_t + D_{t-p+h}, & h = 1, 2, ... \end{cases}$$

Pour la définition de la région d'additivité-inversibilité, l'utilisateur peut se référer à Archibald (1990).

Modèle de Holt-Winters saisonnier multiplicatif

Cette méthode permet de prendre en compte une tendance qui varie avec le temps, et une composante saisonnière de période p. Les prévisions tiennent compte de la tendance et de la saisonnalité. Ce modèle met en jeu trois paramètres. On l'appelle multiplicatif car la composante saisonnière varie avec le temps. Plus les écarts entre les observations sont importants, plus la composante saisonnière augmente. Les équations du modèle sont données par :

$$\begin{cases} Y_t = (\mu_t + \beta_t t)s_p(t) + \epsilon_t \\ P_t Y_{t+h} = (\mu_t + \beta_t t)s_p(t) & h = 1, 2, ... \\ S_t = \alpha(Y_t / S_{t-p}) + (1-\alpha)(S_{t-1} + T_{t-1}) \\ T_t = \beta(S_t - S_{t-1}) + (1-\beta)T_{t-1} \\ D_t = \gamma(Y_t / S_t) + (1-\gamma)D_{t-p} \\ \hat{Y}_{t+h} = P_t Y_{t+h} = (S_t + hT_t)D_{t-p+h}, & h = 1, 2, ... \end{cases}$$

Pour la définition de la région d'additivité-inversibilité, l'utilisateur peut se référer à Archibald (1990).

Remarque 1 : pour les modèles définis ci-dessus, XLSTAT estime les paramètres en cherchant la solution du minimum de la somme du carré des erreurs (SCE). Il est aussi possible de rechercher la solution qui maximise la vraisemblance, sachant qu'en dehors du modèle de Holt-Winters multiplicatif, il est possible d'exprimer les modèles sous la forme d'un modèle ARIMA. Par exemple, le lissage exponentiel simple est équivalent à un modèle ARIMA(0,1,1), le lissage exponentiel double est équivalent à un modèle ARIMA(0,2,2) et le modèle de Holt-Winters additif peut s'écrire sous la forme d'un modèle ARIMA (0,1,p+1)(0,1,0). Si vous préférez maximiser la vraisemblance, nous vous invitons à utiliser la procédure ARIMA de XLSTAT.

Remarque 2 : **pour les modèles ci-dessus, des valeurs initiales pour** S, T et D, sont nécessaires. XLSTAT offre différentes options, y compris du *backcasting*, pour définir les valeurs initiales. Lorsque le *backcasting* est choisi, l'algorithme renverse la série, prend des valeurs initiales correspondant à l'option de base Y(x), puis calcule des estimateurs, qui sont ensuite utilisés comme valeurs initiales sur la série originale. Les options disponibles pour les différents modèles sont définies par:

Lissage exponentiel simple:

$$\begin{cases} Y(1): & S_1 = Y_1 \\ Moyenne(6): & S_1 = \sum_{i=1}^{6} Y_i / 6 \\ Backcasting \\ Optimisé \end{cases}$$

Lissage exponentiel double:

$$\begin{cases} Y(1): & S_1 = Y_1, \quad T_1 = Y_1 \\ Moyenne(6): & S_1 = \sum_{i=1}^{6} Y_i/6, \quad T_1 = S_1 \\ Backcasting & \end{cases}$$

Lissage exponentiel linéaire de Holt :

$$\begin{cases} 0: & S_1 = 0 \\ Backcasting & \end{cases}$$

Modèle de Holt-Winters saisonnier additif :

$$\begin{cases} Y(1+p): & S_{1+p} = \sum_{i=}^{p} Y_i/p, \\ & T_{1+p} = 0 \\ & D_t = Y_t - (Y_1 + T_{1+p}(i-1)) \;\; i = 1, ...p \\ Backcasting & \end{cases}$$

Modèle de Holt-Winters saisonnier multiplicatif :

$$\begin{cases} Y(1+p): & S_{1+p} = \sum_{i=}^{p} Y_i/p, \\ & T_{1+p} = 0 \\ & D_t = Y_t/(Y_1 + T_{1+p}(i-1)) \;\; i = 1, ...p \\ Backcasting & \end{cases}$$

Moyenne mobile

Cette moyenne permet de prendre en compte de manière simple et contrôlée des observations passées pour prédire le futur. Néanmoins l'utilité de la méthode réside plus dans sa nature de filtre, permettant de retirer à une série son bruit de fond, et de faire alors ressortir les grandes tendances. Alors que pour les méthodes précédentes, toute observation a une influence, aussi légère soit-elle, sur les prévisions suivantes, ici, le nombre d'observations du passé prises en compte est limité à q. Les moyennes mobiles servant souvent de filtre, on appelle q la bande passante. Les équations du modèle sont données par :

$$\begin{cases} Y_t = \mu_t + \epsilon_t \\ \hat{\mu_t} = \frac{\sum_{i=q}^{ql} w_i Y_{t+i}}{\sum_{i=q}^{ql} w_i} \end{cases}$$

où l est une constante qui, fixée à zéro, fait en sorte que la prévision dépend des q valeurs passées et de la valeur actuelle. Si l est fixée à un, la prévision dépend aussi des q valeurs suivantes.

où w_i (i=1…q) correspond aux poids des observations autour de Y_t. Les poids peuvent être constants, fixés par l'utilisateur, ou fondés sur des définitions de poids optimaux correspondant

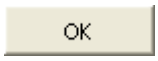
à certains objectifs ; XLSTAT permet l'utilisation de la pondération Spencer 15-point qui laisse passer des polynômes de degré 3.

Lissage de Fourier

Le principe du lissage de Fourier est d'effectuer une transformée de Fourier, et ne retenir qu'une partie du spectre, puis de faire une transformée inverse afin d'obtenir la série lissée.

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.

Onglet **Général**:

Séries temporelles : sélectionnez la ou les séries temporelles que vous voulez analyser. Si des en-têtes de colonnes ont été sélectionnés, veuillez vérifier que l'option « Libellés des séries » est activée.

Données de date : activez cette option pour sélectionner des données date. Ces données doivent être au format de data Excel, ou des valeurs numériques.

- **Vérifier les intervalles** : activez cette option si vous voulez que XLSTAT vérifie que les données de date sont bien régulièrement espacées.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des séries : activez cette option si la première ligne des données sélectionnées (variables dépendantes et explicatives, libellés des observations, poids) contient un libellé.

Modèle : choisissez le modèle de lissage à utiliser (voir description pour plus de détails sur les différents modèles).

Onglet **Options**:

Méthode : choisissez la méthode pour le modèle choisi.

Conditions d'arrêt :

- **Itérations** : entrez le nombre maximal d'itérations pour l'algorithme d'optimisation. Les calculs sont interrompus dès que le nombre maximal d'itérations est dépassé. Valeur par défaut : 500.
- **Convergence** : entrez la valeur seuil d'évolution maximale des communalités d'une itération à l'autre qui, une fois atteinte, permet de considérer que l'algorithme a convergé. Valeur par défaut : 0,00001.

Intervalles de confiance : entrez la valeur de l'intervalle de confiance pour les prédictions effectuées sur l'échantillon de validation et de prédiction.

S1 : choisissez la méthode d'estimation pour les valeurs de départ. Voir la description pour plus de détails.

En fonction du type de modèle et de la méthode choisie, différentes options sont affichées dans la boîte de dialogue. Dans la section description, vous trouverez des informations sur les différents modèles et leurs paramètres.

Dans le cas des modèles exponentiels ou de Holt-Winters, vous pouvez choisir de fixer ou d'optimiser les paramètres. Pour les modèles de Holt-Winters saisonniers vous devez saisir la valeur de la **période**.

Dans le cas du lissage de Fourier, vous devez entrer la proportion **p** du spectre à conserver après le filtrage des hautes fréquences.

Pour la moyenne mobile vous devez spécifier le nombre de pas de temps **q** à utiliser autour de la valeur prédite. Vous pouvez éventuellement ne considérer que la partie **gauche** (valeurs

précédentes uniquement) de la série. L'option « **strictement** » vous permet de ne prendre en compte que les valeurs précédentes.

Onglet **Validation** :

Validation : activez cette option si vous souhaitez utiliser une partie des données sélectionnées pour valider le modèle.

Pas de temps : entrez le nombre de pas de temps à la fin de la série sélectionnée qui doit être utilisé pour valider le modèle choisi.

Onglet **Prédiction** :

Prédiction : activez cette option pour effectuer des prédictions de nouvelles valeurs.

Pas de temps : entrez le nombre de pas de temps à prédire.

Onglet **Données manquantes** :

Ne pas accepter les valeurs manquantes : activez cette option pour que XLSTAT empêche la poursuite des calculs si des valeurs manquantes sont détectées.

Supprimer les observations : activez cette option pour supprimer les observations comportant des données manquantes.

- **Vérifier pour chaque série** : activez cette option pour supprimer les observations comportant des données manquantes série par série.
- **Pour tous les Y** : activez cette option pour supprimer les observations comportant des données manquantes pour l'ensemble des séries.

Remplacer par la moyenne des valeurs précédente et suivante : activez cette option pour estimer les données manquantes par la moyenne de la première valeur précédente non manquante et de la première valeur suivante non manquante.

Ignorer les données manquantes : activez cette option pour ignorer les données manquantes et les estimer au travers du modèle choisi.

Onglet **Sorties**:

Statistiques descriptives : activez cette option pour afficher les statistiques descriptives des séries sélectionnées.

Coefficients d'ajustement : activez cette option pour afficher le tableau des statistiques d'ajustement du modèle.

Paramètres du modèle : activez cette option pour afficher le tableau des paramètres du modèle.

Prédictions et résidus : activez cette option pour afficher les prédictions et les résidus pour l'ensemble des observations.

Onglet **Graphiques** :

Afficher les graphiques : activez cette option pour afficher les graphiques présentant les séries avant et après lissage, ainsi que le diagramme en bâtons des résidus.

Résultats

Coefficients d'ajustement : dans ce tableau sont affichés les coefficients d'ajustement du modèle. Remarque : les coefficients ne sont calculés que sur la base des données utilisées pour l'ajustement et les données de validation ne sont donc pas prises en compte.

Paramètres du modèle : dans ce tableau sont affichés les estimateurs des paramètres du modèle. Remarque : à S1 correspond la première valeur calculée pour la série S, et à T1 correspond la première valeur calculée pour la série T. Voir la description pour plus de détails.

Série avant et après lissage : dans ce tableau sont affichés la série originale et la série lissée ainsi que les résidus et les intervalles de confiance dans le cas où une validation ou des prédictions ont été demandées.

Graphiques : deux graphiques sont affichés. Le premier graphique permet de visualiser les données, le modèle, les prévisions (validation et nouvelles observations) de même que les intervalles de confiance sur les prévisions. Le second graphique permet de visualiser les résidus du modèle.

Exemple

Un exemple de lissage par la méthode de Holt-Winters est disponible sur le Centre d'aide XLSTAT :

http://www.xlstat.com/demo-hwf.htm

Bibliographie

Archibald B.C. (1990). Parameter space of the Holt-Winters' model. *International Journal of Forecasting*, **6**, 199-209.

Box G. E. P. and Jenkins G. M. (1976). Time Series Analysis: Forecasting and control. Holden-Day, San Francisco.

Brockwell P.J. and Davis R.A. (1996). Introduction to Time Series and Forecasting. Springer Verlag, New York.

Brown R.G. (1962). Smoothing, Forecasting and Prediction of Discrete Time Series. Prentice-Hall, New York.

Brown R.G. and Meyer R.F. (1961). The fundamental theorem of exponential smoothing. *Operations Research*, **9**, 673-685.

Chatfield, C. (1978). The Holt-Winters forecasting procedure. *Applied Statistics*, **27**, 264-279.

Holt C.C. (1957). Forecasting seasonals and trends by exponentially weighted moving averages. ONR Reseach Memorandum 52, Carnegie Institute of Technology, Pittsburgh.

Makridakis S.G., Wheelwright S.C. and Hyndman R.J. (1997). Forecasting : Methods and Applications. John Wiley & Sons, New York.

Shumway R.H. and Stoffer D.S. (2000). Time Series Analysis and Its Applications. Springer Verlag, New York.

Winters P.R. (1960). Forecasting sales by exponentially weighted moving averages. *Management Science*, **6**, 324-342

ARIMA

Utilisez cet outil pour ajuster un modèle ARMA (Autoregressive Moving Average), un modèle ARIMA (Autoregressive Integrated Moving Average) ou un modèle SARIMA (Seasonal Autoregressive Integrated Moving Average) ou SARIMAX (avec des variables explicatives) et faire des prévisions sur la base de modèles dont les coefficients sont connus ou à estimer.

Dans cette section :

Description

Les modèles de la famille ARIMA permettent de représenter sous une forme succincte certains phénomènes variant avec le temps, et de faire des prévisions pour les valeurs futures du phénomène, avec un intervalle de confiance autour des prévisions.

L'écriture mathématique des modèles ARIMA varie d'un auteur à l'autre, ceci impliquant notamment des différences pour les signes des coefficients. La notation utilisée dans XLSTAT-Time correspond à celle de la plupart des logiciels.

Soit $\{X_t\}$ une série chronologique de moyenne μ. Si la série suit un modèle ARIMA $(p,d,q)(P,D,Q)^s$, alors on peut écrire :

$$\begin{cases} Y_t = (1-B)^d(1-B^s)^D X_t - \mu \\ \phi(B)\Phi(B^s)Y_t = \theta(B)\Theta(B^s)Z_t,\ Z_t \sim N(0,\sigma^2) \end{cases}$$

avec

$$\begin{cases} \phi(z) = 1 - \sum_{i=1}^{p} \phi_i z^i, \quad \Phi(z) = 1 - \sum_{i=1}^{P} \Phi_i z^i \\ \theta(z) = 1 + \sum_{i=1}^{q} \theta_i z^i, \quad \Theta(z) = 1 + \sum_{i=1}^{Q} \Theta_i z^i \end{cases}$$

p est l'ordre de la partie autorégressive du modèle.

q est l'ordre de la partie moyenne mobile du modèle.

d est l'ordre de différentiation du modèle.

D est l'ordre de différentiation du modèle pour la partie saisonnière.

s est la période du modèle (par exemple 12 si les données sont mensuelles et que l'on a repéré une cyclicité à l'échelle de l'année.

P est l'ordre de la partie autorégressive saisonnière du modèle.

Q est l'ordre de la partie moyenne mobile saisonnière du modèle.

Remarque 1 : le processus $\{Y_t\}$ est causal si et seulement si pour tout z tel que $|z| \leq 1$, $f(z) \neq 0$ et $q(z) \neq 0$.

Remarque 2 : si D=0, on se trouve dans le cas d'un modèle ARIMA(p,d,q). Dans ce cas, P, Q et s sont considérés comme étant nuls.

Remarque 3 : si d=0 et D=0, on se trouve dans le cas d'un modèle ARMA(p,q).

Remarque 4 : si d=0, D=0 et q=0, on se trouve dans le cas d'un modèle AR(p).

Remarque 5 : si d=0, D=0 et p=0, on se trouve dans le cas d'un modèle MA(q).

Si les coefficients des polynômes f, F, q, Q sont inconnus, une fois les paramètres (p,d,q), (P,D,Q) et s saisis, XLSTAT permet d'estimer les coefficients des polynômes, puis de calculer différentes statistiques d'ajustement, et si l'utilisateur le souhaite, de calculer des prévisions de valeurs futures.

Si les coefficients des polynômes f, F, q, Q sont connus, l'utilisateur peut les saisir. XLSTAT calcule ensuite différentes statistiques d'ajustement, et si l'utilisateur le demande, des prévisions de valeurs futures.

Dans le cas où D = 0, il est possible d'effectuer une estimation préliminaire des coefficients des polynômes f et q en utilisant la méthode proposée :

- Si q = 0, deux méthodes d'estimation préliminaire sont proposées. La première utilise l'algorithme de Yule-Walker, le seconde celui de Burg.
- Si p = 0, la méthode utilisée est l'algorithme des innovations.
- Si p = 0 et q = 0, la méthode utilisée est l'algorithme de Hannan-Rissanen.

Dans le cas où D > 0, XLSTAT effectue lui-même la recherche d'un point de départ raisonnable.

Variables explicatives

XLSTAT permet d'inclure des variables explicatives dans le modèle ARIMA. Trois approches sont possibles:

1. OLS : un modèle de régression linéaire classique est ajusté aux données puis les résidus sont modélisés au travers d'un modèle (S)ARIMA.
2. CO-LS : si d ou D et s ne sont pas nuls, les données (y compris les variables explicatives) sont différenciées, puis un modèle ARMA est ajusté en même temps que les coefficients

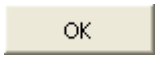
du modèle linéaire sont ajuste suivant la procédure de Cochrane et Orcutt (1949).

3. GLS : Un modèle de régression linéaire classique est ajusté, puis les résidus sont modélisés au travers d'un modèle (S)ARIMA model, puis on revient à l'étape de régression en modifiant les coefficients du modèle linéaire avec un algorithme de Newton Raphson dans le but d'améliorer la vraisemblance.

Remarque : si aucune différenciation n'intervient dans le modèle (d=0 et D=0), et s'il n'y a pas de variable explicative dans le modèle, la constante du modèle est estimée avec l'approche CO-LS.

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.

Onglet **Général**:

Séries temporelles : sélectionnez la ou les séries temporelles que vous voulez analyser. Si des en-têtes de colonnes ont été sélectionnés, veuillez vérifier que l'option « Libellés des séries » est activée.

Centrer : activez cette option pour centrer les séries après différenciation.

Variance : activez cette option puis entrez la valeur de la variance si vous souhaitez imposer une variance des erreurs pour le modèle.

Données de date : activez cette option pour sélectionner des données de date. Ces données doivent être au format de data Excel, ou des valeurs numériques.

- **Vérifier les intervalles** : activez cette option si vous voulez que XLSTAT vérifie que les données de date sont bien régulièrement espacées.

X / Variables explicatives : activez cette option si vous voulez inclure une ou plusieurs variables explicatives quantitatives dans le modèle. Sélectionnez alors la ou les variables correspondantes sur la feuille Excel. Les données sélectionnées doivent être de type numérique. Si le libellé des variables a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

- **Mode** : choisissez la façon dont doivent être traitées les variables explicatives. Les trois modes sont décrits dans la section description.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des séries : activez cette option si la première ligne des données sélectionnées (variables dépendantes et explicatives, libellés des observations, poids) contient un libellé.

Paramètres du modèle : entrez la valeur des différents ordres intervenant dans le modèle :

- **p** : entrez l'ordre de la partie autorégressive du modèle. Par exemple, entrez 1 pour un modèle AR(**1**) ou pour un modèle ARMA(**1**,2).
- **d** : entrez l'ordre de différentiation du modèle. Par exemple, entrez 1 pour un modèle ARIMA(0,**1**,2).
- **q** : entrez l'ordre de la partie moyenne mobile du modèle. Par exemple, entrez 2 pour un modèle MA(**2**) ou pour un modèle ARIMA(1,1,**2**).
- **P** : entrez l'ordre de la partie autorégressive saisonnière du modèle. Par exemple, entrez 1 pour un modèle ARIMA(1,1,0)(**1**,1,0)^{12}. Vous ne pouvez modifier cette valeur que si $D \neq 0$. Si $D = 0$, on considère que $P = 0$.
- **D** : entrez l'ordre de différentiation du modèle pour la partie saisonnière. Par exemple, entrez 1 pour un modèle ARIMA(0,1,1)(0,**1**,1)^{12}.
- **Q** : entrez l'ordre de la partie moyenne mobile saisonnière du modèle. Par exemple, entrez 1 pour un modèle ARIMA(0,1,1)(0,1,**1**)^{12}. Vous ne pouvez modifier cette valeur que si $D \neq 0$. Si $D = 0$, on considère que $Q = 0$.

- **s** : entrez la période du modèle. Vous ne pouvez modifier cette valeur que si $D \neq 0$. Si $D = 0$, on considère que $s = 0$.

Onglet **Options**:

Estimation préliminaire : activez cette option si vous souhaitez utiliser une méthode d'ajustement préliminaire. Cette option n'est disponible que si D=0.

- **Yule-Walker** : activez cette option pour estimer les coefficients du modèle autorégressif AR(p) avec l'algorithme de Yule-Walker.
- **Burg** : activez cette option pour estimer les coefficients du modèle autorégressif AR(p) avec l'algorithme de Burg.
- **Innovations** : activez cette option pour estimer les coefficients du modèle moyenne mobile MA(q) avec l'algorithme des Innovations.
- **Hannan-Rissanen** : activez cette option pour estimer les coefficients du modèle ARMA(p,q) avec l'algorithme de Hannan-Rissanen.
- **m/Auto** : si vous choisissez la méthode des Innovations ou de Hannan-Rissanen, vous devez entrez la valeur m spécifique de chacun des algorithmes. Si vous choisissez Auto, XLSTAT détermine automatiquement quelle est la bonne valeur de m.

Coefficients initiaux : activez cette option pour sélectionner des valeurs initiales des coefficients du modèle.

- **Phi** : sélectionnez à ce niveau la valeur des coefficients correspondant à la partie autorégressive du modèle (y compris pour la partie saisonnière). Le nombre de valeurs sélectionné ici doit être égal à $p + P$.
- **Theta** : sélectionnez à ce niveau la valeur des coefficients correspondant à la partie moyenne mobile du modèle (y compris pour la partie saisonnière). Le nombre de valeurs sélectionné ici doit être égal à $q + Q$.

Optimiser : activez cette option pour estimer les coefficients selon l'une des deux méthodes proposées :

- **Vraisemblance** : activez cette option pour maximiser la vraisemblance.
- **Moindres carrés** : activez cette option pour minimiser la somme des carrés des erreurs.

Conditions d'arrêt :

- **Itérations** : entrez le nombre maximal d'itérations pour l'algorithme d'optimisation. Les calculs sont interrompus dès que le nombre maximal d'itérations est dépassé. Valeur par défaut : 500.

- **Convergence** : entrez la valeur seuil d'évolution maximale des communalités d'une itération à l'autre qui, une fois atteinte, permet de considérer que l'algorithme a convergé. Valeur par défaut : 0,00001.

Rechercher le meilleur modèle : Activez cette option pour explorer combinaisons d'ordres p/q/P/Q. Si vous activez cette option, les ordres minimaux sont ceux entrés dans l'onglet "Général", et les ordres maximaux doivent être définis en utilisant les options suivantes :

- **Max(p)** : entrez la valeur maximum de p à explorer.
- **Max(q)** : entrez la valeur maximum de q à explorer.
- **Max(P)** : entrez la valeur maximum de P à explorer.
- **Max(Q)** : entrez la valeur maximum de Q à explorer.
- **AICC** : activez cette option pour utiliser le AICC (Akaike Information Criterion Corrected) pour identifier le meilleur modèle.
- **SBC** : activez cette option pour utiliser le SBC (Schwarz's Bayesian Criterion) pour identifier le meilleur modèle.

Onglet **Validation** :

Validation : activez cette option si vous souhaitez utiliser une partie des données sélectionnées pour valider le modèle.

Pas de temps : entrez le nombre de pas de temps à la fin de la série sélectionnée qui doit être utilisé pour valider le modèle choisi.

Onglet **Prédiction** :

Prédiction : activez cette option pour effectuer des prédictions de nouvelles valeurs.

Pas de temps : entrez le nombre de pas de temps à prédire.

Onglet **Données manquantes** :

Ne pas accepter les valeurs manquantes : activez cette option pour que XLSTAT empêche la poursuite des calculs si des valeurs manquantes sont détectées.

Supprimer les observations : activez cette option pour supprimer les observations comportant des données manquantes.

- **Vérifier pour chaque série** : activez cette option pour supprimer les observations comportant des données manquantes série par série.

- **Pour tous les Y** : activez cette option pour supprimer les observations comportant des données manquantes pour l'ensemble des séries.

Remplacer par la moyenne des valeurs précédente et suivante : activez cette option pour estimer les données manquantes par la moyenne de la première valeur précédente non manquante et de la première valeur suivante non manquante.

Onglet **Sorties**:

Statistiques descriptives : activez cette option pour afficher les statistiques descriptives des séries sélectionnées.

Coefficients d'ajustement : activez cette option pour afficher le tableau des statistiques d'ajustement du modèle.

Paramètres du modèle : activez cette option pour afficher le tableau des paramètres du modèle.

Prédictions et résidus : activez cette option pour afficher les prédictions et les résidus pour l'ensemble des observations.

Intervalles de confiance : entrez la valeur de l'intervalle de confiance pour les prédictions effectuées sur l'échantillon de validation et de prédiction.

Onglet **Graphiques** :

Afficher les graphiques : activez cette option pour afficher le graphique présentant les données originales et les prédictions du modèle, ainsi que le diagramme en bâtons des résidus.

Résultats

Statistiques simples : tableau dans lequel sont affichés le nombre d'observations, le nombre d'observations manquantes, le minimum, le maximum, la moyenne, la variance de la population (1/n) et l'écart type (1/n).

Si une estimation préliminaire et une optimisation ont été demandées, les résultats de l'estimation préliminaire sont affichés, suivis de ceux de l'optimisation. Si des coefficients initiaux ont été saisis, les résultats concernant ces coefficients sont d'abord affichés.

Coefficients d' ajustement :

- **Observations** : le nombre de données utilisées pour l'ajustement.
- **SCE** : la somme des carrés des résidus. Ce critère est minimisé lorsque l'option « Moindres carrés » est sélectionnée.
- **Variance du bruit blanc** : cette statistique est égale à SCE divisé par N. Dans certains logiciels cette statistique est désignée par sigma2.

- **Variance du bruit blanc (estimée)** : cette statistique est en principe égale à la précédente le diviseur étant DDL, le nombre de degrés de liberté. Dans le cas des algorithmes de Yule-Walker et de Burg, une estimation légèrement différente est fournie.
- **-2Log(Vrais.)** : ce critère est minimisé dans le cas d'une optimisation basée sur le maximum de vraisemblance. Elle vaut l'opposé de deux fois le logarithme népérien de la vraisemblance.
- **FPE** : ce critère est dû à Akaike (Final Prediction Error). Ce critère est adapté pour les modèles autorégressifs.
- **AIC** : ce critère est dû à Akaike (Akaike Information Criterion).
- **AICC** : ce critère est dû à Brockwell (Akaike Information Criterion Corrected).
- **SBC** : ce critère est dû à Schwarz (Schwarz's Bayesian Criterion).

Paramètres du modèle :

Le premier tableau de paramètres correspond aux coefficients du modèle linéaire. Si aucune variable explicative n'a été introduite du modèle, seules les informations concernant la constante sont affichées.

Le tableau suivant donne l'estimateur de chaque coefficient de chaque polynôme, ainsi que l'écart-type obtenu soit directement par la méthode d'estimation (estimation préliminaire) soit à partir de la matrice d'information de Fisher à l'issue de l'optimisation (désignée par Hess., pour Hessienne). Les écarts-types asymptotiques sont aussi calculés. Pour chaque coefficient et chaque écart-type est fourni un intervalle de confiance. Les coefficients sont identifiés de la manière suivante :

AR(i) : coefficient correspondant au coefficient d'ordre i du polynôme f (z).

SAR(i) : coefficient correspondant au coefficient d'ordre i du polynôme F (z).

MA(i) : coefficient correspondant au coefficient d'ordre i du polynôme q (z).

SMA(i) : coefficient correspondant au coefficient d'ordre i du polynôme Q (z).

Prédictions et résidus : dans ce tableau sont affichés la série de départ, les prédictions calculées à partir du modèle, et les résidus correspondants. Si l'utilisateur l'a demandé, des prédictions pour les données de validation et pour les valeurs futures sont calculées, ainsi que les écart-types et les intervalles de confiance correspondants.

Graphiques : deux graphiques sont affichés. Le premier graphique permet de visualiser les données, les valeurs calculées à partir du modèle, les prévisions de validation et des valeurs futures, de même que les intervalles de confiance. Le second graphique permet de visualiser les résidus du modèle.

Exemple

Un exemple d'utilisation de la méthode ARIMA est disponible sur le Centre d'aide XLSTAT :

http://www.xlstat.com/demo-arimaf.htm

Bibliographie

Box G. E. P. and Jenkins G. M. (1984). Time Series Analysis: Forecasting and Control, 3rd edition. Pearson Education, Upper Saddle River.

Brockwell P.J. and Davis R.A. (2002). Introduction to Time Series and Forecasting, 2-nd edition. Springer Verlag, New York.

Brockwell P. J. and Davis R. A. (1991). Time series: Theory and Methods, 2-nd edition. Springer Verlag, New York.

Cochrane D. and Orcutt G.H. (1949). Application of least squares regression to relationships containing autocorrelated error terms. *Journal of the American Statistical Association*, **44**, 32-61.

Fuller W.A. (1996). Introduction to Statistical Time Series, Second Edition. John Wiley & Sons, New York.

Hannan E.J. and Rissanen J. (1982). Recursive estimation of mixed autoregressive-moving average models order. *Biometrika*, **69**, 1, 81-94.

Mélard G. (1984). Algorithm AS197: a fast algorithm for the exact likelihood of autoregressive-moving average models. *Journal of the Royal Statistical Society, Series C, Applied Statistics*, **33**, 104-114.

Percival D. P. and Walden A. T. (1998). Spectral Analysis for Physical Applications. Cambridge University Press, Cambridge.

Analyse spectrale

Utilisez cet outil pour transformer une série chronologique en ses coordonnées dans l'espace des fréquences, et pour analyser ses caractéristiques dans le nouvel espace.

Dans cette section :

Description

La représentation spectrale d'une série chronologique $\{X_t\}$, $(t = 1, \ldots, n)$, consiste en la décomposition de $\{X_t\}$ en une sommes de sinusoïdes avec des coefficients aléatoires non corrélés. On peut en déduire une décomposition des fonctions de variance et d'autocovariance en une somme de sinusoïdes.

La densité spectrale correspond en théorie à la décomposition d'une série chronologique. Cependant, dans la pratique, on n'a accès qu'à un nombre limité de données, échantillonnées en général à intervalles réguliers de temps. Pour cette raison, on doit dans un premier temps obtenir les coordonnées de la transformée de Fourier (partie réelle et partie imaginaire), puis le périodogramme, à partir duquel, grâce à une méthode de lissage on pourra obtenir une estimation de la densité spectrale qui est une meilleure représentation du spectre.

XLSTAT calcule automatiquement les parties réelles et imaginaires pour chaque fréquence en s'appuyant sur des méthodes rapides et performantes de calcul de la transformation de Fourier, puis calcule et affiche les résultats présentés ci-dessous.

Si n est la taille de l'échantillon, et si $[i]$ désigne l'entier le plus grand plus petit ou égal à i, alors les fréquences de Fourier sont données par :

$$\omega_k = \frac{2\pi k}{n}, k = -[\frac{n-1}{2}], \ldots [\frac{n}{2}]$$

Les composantes cosinus et sinus de la transformée de Fourier sont données par :

$$a_k = \frac{2}{n} \sum_{t=1}^{n} X_t \cos(\omega_k (t-1))$$

$$b_k = \frac{2}{n} \sum_{t=1}^{n} X_t \sin(\omega_k (t-1))$$

Le périodogramme est donné par :

$$I_k = \frac{n}{2}\sum_{t=1}^{n}(a_k^2 + b_k^2)$$

L'estimateur de la densité spectrale de la série chronologique $\{X_t\}$ est donné par :

$$\hat{f}_k = \sum_{i=-p}^{p} w_i J_{k+i}$$

avec

$$\begin{cases} J_{k+i} = I_{k+i}, & 0 \leq k+i \leq n \\ J_{k+i} = I_{-(k+i)}, & k+i < 0 \\ J_{k+i} = I_{n-(k+i)}, & k+i > n \end{cases}$$

où p, la bande passante, et w_i , les poids, sont soit fixés par l'utilisateur, soit déterminés par le choix d'un noyau.

Si on définit, $p = c.q^e$, $q = [n/2] + 1$, et $\lambda_i = i/p$, les noyaux proposés par XLSTAT sont :

Bartlett :

$$\begin{cases} c = 1/2, & e = 1/3 \\ w_i = 1 - |\lambda_i| & \text{si } |\lambda_i| \leq 1 \\ w_i = 0 & \text{sinon} \end{cases}$$

Parzen :

$$\begin{cases} c = 1, & e = 1/5 \\ w_i = 1 - 6|\lambda_i|^2 + 6|\lambda_i|^3 & \text{si } |\lambda_i| \leq 0.5 \\ w_i = 2(1 - |\lambda_i|)^3 & \text{si } 0.5 \leq |\lambda_i| \leq 1 \\ w_i = 0 & \text{sinon} \end{cases}$$

Quadratic spectral :

$$\begin{cases} c = 1/2, e = 1/5 \\ w_i = \frac{25}{12\pi^2\lambda_i^2}\left(\frac{\sin(6\pi\lambda_i/5)}{6\pi\lambda_i/5} - \cos(6\pi\lambda_i/5)\right) \end{cases}$$

Tukey-Hanning :

$$\begin{cases} c = 2/3, & e = 1/5 \\ w_i = (1 + \cos(\pi\lambda_i))/2 & \text{si } |\lambda_i| \leq 1 \\ w_i = 0 & \text{sinon} \end{cases}$$

Tronqué :

$$\begin{cases} c = 1/4, & e = 1/5 \\ w_i = 1 & \text{si } |\lambda_i| \leq 1 \\ w_i = 0 & \text{sinon} \end{cases}$$

Remarque : la bande passante p est une fonction de n, la taille de l'échantillon. Les poids w_i doivent être positifs et avoir pour somme 1. Si tel n'est pas le cas, XLSTAT les normalise automatiquement.

Si une seconde série est disponible, plusieurs fonctions supplémentaires peuvent être calculées pour estimer le spectre croisé.

La partie réelle du périodogramme croisé de $\{X_t\}$ et $\{Y_t\}$ est donnée par :

$$Real_k = \frac{n}{2} \sum_{t=1}^{n} (a_{X,k} a_{Y,k} + b_{X,k} b_{Y,k})$$

La partie imaginaire du périodogramme croisé de $\{X_t\}$ et $\{Y_t\}$ est donnée par :

$$Imag_k = \frac{n}{2} \sum_{t=1}^{n} (a_{X,k} b_{Y,k} - b_{X,k} a_{Y,k})$$

L'estimation du cospectre (partie réelle du spectre croisé) des séries $\{X_t\}$ et $\{Y_t\}$ est donné par :

$$C_k = \sum_{i=-p}^{p} w_i R_{k+i}$$

avec

$$\begin{cases} R_{k+i} = Real_{k+i}, & 0 \leq k+i \leq n \\ R_{k+i} = Real_{-(k+i)}, & k+i < 0 \\ R_{k+i} = Real_{n-(k+i)}, & k+i > n \end{cases}$$

L'estimation du spectre quadratique (partie imaginaire du spectre croisé) des séries $\{X_t\}$ et $\{Y_t\}$ est donnée par :

$$Q_k = \sum_{i=-p}^{p} w_i H_{k+i}$$

avec

$$\begin{cases} H_{k+i} = Imag_{k+i}, & 0 \leq k+i \leq n \\ H_{k+i} = Imag_{-(k+i)}, & k+i < 0 \\ H_{k+i} = Imag_{n-(k+i)}, & k+i > n \end{cases}$$

La phase du spectre croisé de $\{X_t\}$ et $\{Y_t\}$ est donnée par :

$$\Phi_k = \arctan(Q_k/C_k)$$

L'amplitude de spectre croisé de $\{X_t\}$ et $\{Y_t\}$ est donnée par :

$$A_k = \sqrt{C_k^2 + Q_k^2}$$

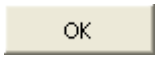
La cohérence carrée des séries $\{X_t\}$ et $\{Y_t\}$ est donnée par :

$$K_k = \frac{A_k^2}{\hat{f}_{X,k}\hat{f}_{Y,k}}$$

Tests du bruit blanc : XLSTAT vous propose en option de calculer deux statistiques et la p-value associée, afin de déterminer si la série est significativement différente d'un bruit blanc ou non : le Kappa de Fisher et la statistique du Kolmogorov-Smirnov de Bartlett.

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.

Onglet **Général** :

Séries temporelles : sélectionnez la ou les séries temporelles dont vous voulez analyser le spectre. Si des en-têtes de colonnes ont été sélectionnés, veuillez vérifier que l'option « Libellés des séries » est activée.

Données de date : activez cette option pour sélectionner des données date. Ces données doivent être au format de data Excel, ou des valeurs numériques.

Vérifier les intervalles : activez cette option si vous voulez que XLSTAT vérifie que les données de date sont bien régulièrement espacées.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des séries : activez cette option si la première ligne des données sélectionnées (variables dépendantes et explicatives, libellés des observations, poids) contient un libellé.

Onglet **Données manquantes** :

Ne pas accepter les valeurs manquantes : activez cette option pour que XLSTAT empêche la poursuite des calculs si des valeurs manquantes sont détectées.

Supprimer les observations : activez cette option pour supprimer les observations comportant des données manquantes.

Remplacer par la moyenne des valeurs précédente et suivante : activez cette option pour estimer les données manquantes par la moyenne de la première valeur précédente non manquante et de la première valeur suivante non manquante.

Onglet **Sorties (1)** :

Tests de bruit blanc : activez cette option pour afficher les résultats des tests de bruit blanc.

Partie cosinus : activez cette option pour afficher la partie réelle de la transformée de Fourier.

Partie sinus : activez cette option pour afficher la partie imaginaire de la transformée de Fourier.

Amplitude : activez cette option pour afficher l'amplitude du spectre.

Phase : activez cette option pour afficher la phase du spectre.

Densité spectrale : activez cette option pour afficher une estimation de la densité spectrale. Deux options vous sont proposées :

- **Pondération par noyau** : choisissez alors la fonction noyau à utiliser (voir description).
- **c** : entrez la valeur du paramètre c. Ce paramètre est décrit dans la partie description.

- **e** : entrez la valeur du paramètre e. Ce paramètre est décrit dans la partie description.

- **Pondération fixe** : sélectionnez sur une feuille Excel les données correspondant aux poids utilisés pour le lissage. Le nombre de poids doit être impair. L'utilisation de poids symétriques est recommandée (exemple : 1,2,3,2,1).

Onglet **Sorties (2)** :

Spectre croisé : activez cette option pour faire l'analyse des spectres croisés. Ces calculs ne sont effectués que si au moins deux séries temporelles ont été sélectionnées.

- **Partie réelle** : activez cette option pour afficher la partie réelle du spectre croisé.

- **Partie imaginaire** : activez cette option pour afficher la partie imaginaire du spectre croisé.

- **Cospectre** : activez cette option pour afficher le cospectre.

- **Spectre de quadrature** : activez cette option pour afficher le spectre de quadrature.

- **Cohérence carrée** : activez cette option pour afficher la cohérence carrée.

Onglet **Graphiques** :

Périodogramme : activez cette option pour afficher le périodogramme des séries.

Densité spectrale : activez cette option pour afficher le graphique des densités spectrales.

Résultats

Tests du bruit blanc : vous trouverez dans ce tableau pour chaque série, le Kappa de Fisher et la statistique du Kolmogorov-Smirnov de Bartlett ainsi que les p-values correspondantes. Si les p-values sont inférieures au niveau de signification que vous vous êtes fixé (typiquement 0.05), alors vous devez rejeter l'hypothèse que les séries sont un simple bruit blanc.

Analyse spectrale : ce tableau est affiché pour toutes les séries sélectionnées. Les résultats affichés correspondent aux différentes options de sortie sélectionnées. Le **périodo gramme**, qui correspond à l'amplitude du spectre croisé, et le graphique de la **densité spectrale** sont affichés à la suite du tableau.

Analyse du spectre croisé : ce tableau est affiché pour tous les couples de séries sélectionnés.

Exemple

Un exemple d'analyse spectrale est disponible en permanence sur le Centre d'aide XLSTAT :

http://www.xlstat.com/demo-spectralf.htm

Bibliographie

Bartlett M.S. (1966). An Introduction to Stochastic Processes, Second Edition. Cambridge University Press, Cambridge.

Brockwell P.J. and Davis R.A. (1996). Introduction to Time Series and Forecasting. Springer Verlag, New York.

Davis H.T. (1941). The Analysis of Economic Time Series. Principia Press, Bloomington.

Durbin J. (1967). Tests of Serial Independence Based on the Cumulated Periodogram. *Bulletin of Int. Stat. Inst.*, **42**, 1039-1049.

Chiu S-T (1989). Detecting periodic components in a white Gaussian time series. *Journal of the Royal Statistical Society, Series B*, **51**, 249-260.

Fuller W.A. (1996). Introduction to Statistical Time Series, Second Edition. John Wiley & Sons, New York.

Nussbaumer H.J. (1982). Fast Fourier Transform and Convolution Algorithms, Second Edition. Springer-Verlag, New York.

Parzen E. (1957). On Consistent Estimates of the Spectrum of a Stationary Time Series. *Annals of Mathematical Statistics*, **28**, 329-348.

Shumway R.H. and Stoffer D.S. (2000). Time Series Analysis and Its Applications. Springer Verlag, New York.

Transformée de Fourier

Utilisez cet outil pour transformer une série chronologique (ou un signal) en ses coordonnées dans l'espace des fréquences, ou pour effectuer l'opération inverse.

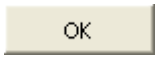
Dans cette section :

Description

Utilisez cet outil pour transformer une série chronologique (ou un signal) en ses coordonnées dans l'espace des fréquences, ou pour effectuer l'opération inverse. Alors que la fonction équivalente de Excel vous limite à des tailles d'échantillon en puissance de 2, XLSTAT accepte une taille quelconque de signal.

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.

Onglet **Général**:

Partie réelle : activez cette option pour sélectionner le signal à transformer, ou la partie réelle dans le cas d'une transformation inverse.

Partie imaginaire : activez cette option pour sélectionner la partie imaginaire dans le cas d'une transformation inverse.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des colonnes : activez cette option si la première ligne des données sélectionnées (partie réelle, partie imaginaire) contient un libellé.

Transformée inverse : activez cette option pour calculer l'inverse de la transformée de Fourier.

Amplitude : activez cette option pour afficher l'amplitude du spectre.

Phase : activez cette option pour afficher la phase du spectre.

Onglet **Données manquantes** :

Ne pas accepter les valeurs manquantes : activez cette option pour que XLSTAT empêche la poursuite des calculs si des valeurs manquantes sont détectées.

Supprimer les observations : activez cette option pour supprimer les observations comportant des données manquantes.

Résultats

Partie réelle : partie réelle obtenue après la transformée ou la transformée inverse.

Partie imaginaire : partie imaginaire obtenue après la transformée ou la transformée inverse.

Amplitude : amplitude du spectre.

Phase : phase du spectre.

Bibliographie

Fuller W.A. (1996). Introduction to Statistical Time Series, Second Edition. John Wiley & Sons, New York.

Simulations Monte Carlo

XLSTAT-Sim

XLSTAT-Sim est un outil à la fois puissant et convivial permettant de créer des modèles utilisant des simulations Monte Carlo afin d'estimer la distribution de variables complexes.

Dans cette section :

Introduction à XLSTAT-Sim

XLSTAT-Sim est un module qui permet de construire et de calculer des modèles de simulation, une méthode novatrice pour l'estimation de variables dont la valeur exacte n'est pas connue, mais qui peuvent être estimées au moyen de simulations de variables aléatoires qui suivent certaines lois de distribution. Avant de lancer le modèle, vous devez créer le modèle en définissant une série de variables d'entrée et de sortie.

Modèles de simulation

Les modèles de simulation permettent d'obtenir des informations, telles que la moyenne ou médiane, pour des variables qui n'ont pas une valeur exacte, mais pour lesquelles nous pouvons connaître, supposer ou calculer une distribution. Si des variables « résultat » dépendent de ces variables « distribution » au travers d'une formule établie, elles auront par voie de conséquence aussi une distribution et non une valeur fixe. XLSTAT-Sim vous permet de définir les distributions, puis d'obtenir, par le biais de simulations itératives après convergence du modèle, une distribution empirique pour les variables d'entrée et de sortie ainsi que les statistiques correspondantes.

Les modèles de simulation sont utilisés dans de nombreux domaines tels que la finance et l'assurance, la médecine, la prospection pétrolière et minière, ou la prévision des ventes.

Quatre types d'objets sont nécessaires pour la construction d'un modèle de simulation :

- Distributions : cet objet correspond à une variable aléatoire dont on choisit la distribution parmi un choix de plus de 20 distributions proposées par XLSTAT, afin d'exprimer l'incertitude quant aux valeurs que peut prendre la variable aléatoire (voir le chapitre Définir une distribution pour plus de détails). Par exemple, on choisira une distribution triangulaire lorsque l'on a une quantité que l'on sait pouvoir varier entre deux bornes mais avec une valeur qui semble plus probable. A chaque itération du calcul du modèle de simulation, un tirage aléatoire est effectué dans chacune des distributions.

- Variables scénario : elles permettent d'introduire dans le modèle de simulation une quantité fixe pour un modèle de simulation donné, mais que l'on fait varier entre deux bornes avec un pas donné, afin d'étudier la sensibilité des variables résultats à ces variables. Autrement dit, on recalcule le modèle de simulation pour chacune des valeurs des variables scénario. Facultatives, les variables de décision sont néanmoins nécessaires pour les graphiques tornado (voir le chapitre Définir une variable scénario pour plus de détails).

- Variables résultat : les variables résultat sont des quantités qui dépendent directement ou indirectement, au travers de formules Excel, des variables aléatoires auxquelles ont été affectées des distributions, et éventuellement des variables de décision. Le but des calculs d'un modèle de simulation est justement de connaître la distribution des variables résultats (voir le chapitre Définir une variable résultat pour plus de détails).

- Statistiques : on peut définir une statistique associée à une distribution, à une variable résultat, ou à une autre statistique. Elle est calculée à chaque itération du calcul du modèle de simulation. Le rapport de simulation inclut alors des résultats concernant la statistique définie. Un grand nombre de statistiques est proposé par XLSTAT (voir le chapitre Définir une statistique pour plus de détails).

Un modèle doit comprendre au moins une distribution et une variable résultat, et autant des quatre objets définis ci-dessus que vous le souhaitez.

Un modèle peut être limité à une unique feuille Excel ou peut utiliser tout un classeur.

Les modèles de simulation permettent de tenir compte des dépendances entre les variables aléatoires décrites par des distributions. Si vous savez que deux variables sont généralement liées avec un coefficient de corrélation entre elles de 0.4, alors vous voulez, quand vous faites des simulations que les échantillons obtenus pour ces deux variables au cours des tirages aléatoires vérifient la même propriété. XLSTAT-Sim rend cela possible : il suffit d'entrer dans la boîte de dialogue permettant de lancer les simulations, la matrice de corrélation ou de covariance entre certaines ou toutes des variables aléatoires utilisées dans le modèle.

Sorties

Lorsque vous lancez les calculs d'un modèle, une série de résultats est affichée par XLSTAT. Tout en fournissant des statistiques essentielles notamment sur les variables d'entrée et de sortie, le rapport de résultats permet aussi d'interpréter la relation entre les différentes variables et, si des variables de décision ont été incluses dans le modèle, d'effectuer une analyse de sensibilité.

Statistiques descriptives :

Le rapport qui est généré après les simulations contient des informations sur les distributions du modèle. L'utilisateur peut choisir parmi un grand nombre de statistiques descriptives les indicateurs les plus importants qui doivent être inclus dans le rapport afin d'interpréter facilement les résultats. Des graphiques sont également disponibles pour permettre de visualiser la distribution des variables et les relations entre elles.

Les détails et les formules relatifs aux statistiques descriptives sont disponibles dans le chapitre consacré à l'outil "Statistiques descriptives" de XLSTAT.

Graphiques :

Les graphiques suivants sont disponibles pour afficher des informations sur les variables:

- **Box plots** : ces représentations univariées d'échantillons de données quantitatives sont parfois appelées « diagrammes boîtes et moustaches ». C'est une représentation simple et assez complète puisque dans la version proposée par XLSTAT sont affichés le minimum, le 1-er quartile, la médiane, la moyenne, le 3-ième quartile, ainsi que les deux limites (les extrémités des « moustaches ») au-delà desquelles on peut considérer que les valeurs sont anormales. La moyenne est affichée sous la forme d'un + rouge, et la médiane sous la forme d'une ligne noire. Les limites sont ainsi calculées :

Limite inférieure : $L_{inf} = X(i)$ tel que $\{X(i) - [Q1 - 1.5(Q3 - Q1)]\}$ soit minimal et $X(i) = Q1 - 1.5(Q3 - Q1)$.

Limite supérieure : $L_{sup} = X(i)$ tel que $\{X(i) - [Q3 + 1.5(Q3 - Q1)]\}$ soit minimal et $X(i) = Q3 + 1.5(Q3 - Q1)$.

Les valeurs en dehors de l'intervalle $]Q1 - 3(Q3 - Q1); Q3 + 3(Q3 - Q1)[$ sont affichées avec un symbole *; et les valeurs comprises dans $[Q1 - 3(Q3 - Q1); Q1 - 1.5(Q3 - Q1)]$ ou $[Q3 + 1.5(Q3 - Q1); Q3 + 3(Q3 - Q1)]$ sont affichées avec un symbole "o".

- **Scattergrams** : ces représentations univariées permettent de donner une idée de la distribution et de la pluralité éventuelle des modes d'un échantillon. Tous les points sont représentés, ainsi que la moyenne et la médiane.

- **Graphiques P-P (loi normale)** : les graphiques Probabilité-Probabilité (*P-P plots* en anglais) permettent de comparer la fonction de répartition empirique d'un échantillon à celle d'un échantillon distribué suivant une loi normale de même moyenne et même variance. Si l'échantillon suit une loi normale, les points doivent être confondus avec la première bissectrice du plan.

- **Graphiques Q-Q (loi normale)** : les graphiques Quantile-Quantile (*Q-Q plots* en anglais) permettent de comparer les quantiles de l'échantillon à ceux d'un échantillon distribué suivant une loi normale de même moyenne et même variance. Si l'échantillon suit une loi normale, les points doivent être confondus avec la première bissectrice du plan.

Corrélations :

Une fois les calculs terminés, le rapport de simulation peut contenir, si l'utilisateur les ad demandées, des informations sur les corrélations entre les différentes variables incluses dans le modèle de simulation. Trois types de coefficients de corrélation sont proposés :

Le coefficient de corrélation de Pearson : cette statistique est le coefficient de corrélation le plus communément utilisé car bien adapté aux données quantitatives continues. Sa valeur est comprise entre -1 et 1, et il mesure le niveau de relation linéaire entre deux variables. Remarque : le coefficient de Pearson au carré, appelé R^2, donne une idée de la proportion de variabilité d'une variable explicable par l'autre. Les p-values calculées pour les coefficients de corrélation permettent de tester l'hypothèse nulle de corrélation non significativement différente de zéro entre les variables. Cependant, il convient d'être prudent car, si l'indépendance entre deux variables implique la nullité du coefficient de corrélation entre les variables, la réciproque n'est pas vraie : on peut avoir une corrélation proche de zéro entre deux variables parce que la relation n'est pas linéaire, ou parce qu'elle est complexe et nécessite la prise en compte d'autres variables.

Le coefficient de corrélation de Spearman (rho) : ce coefficient utilise les rangs des observations et non leur valeur en tant que telle. Ce coefficient est donc adapté aux données ordinales. Comme pour le coefficient de Pearson, on peut aussi interpréter ce coefficient en termes de variabilité expliquée. Ici, il s'agit bien entendu de la variabilité des rangs.

Le coefficient de corrélation de Kendall (tau) : comme pour le coefficient de Spearman, ce coefficient est bien adapté aux variables ordinales car aussi basé sur les rangs. Il est cependant conceptuellement très différent. Il peut être interprété comme en termes de probabilité : c'est la différence entre la probabilité pour que les variables varient dans le même sens et la probabilité pour qu'elles varient dans le sens contraire. Lorsque le nombre d'observations est inférieur à 50 et qu'il n'y a pas d'ex-æquo, XLSTAT fournit la p-value exacte. Sinon une approximation est utilisée. Cette dernière est réputée fiable, dès lors qu'il y a plus de 8 observations.

Analyse de sensibilité :

L'analyse de sensibilité donne des informations sur l'impact des différentes variables d'entrée (variables aléatoires et variables scénario) pour une variable résultat. Sur la base des résultats de la simulation et en fonction du type de corrélation choisi (voir ci-dessus), les corrélations entre les variables aléatoires et de résultat variables sont calculés et affichés dans un ordre décroissant d'impact sur la variable résultat.

Analyse tornado et graphique araignée :

Les analyses tornado et araignée ne sont pas fondées sur les itérations de la simulation, mais sur une analyse point par point de l'ensemble des variables d'entrée (variables aléatoires et variables scénario).

Au cours de l'analyse tornado, pour chaque variable résultat, les variables aléatoires et les variables scénario sont étudiées une par une. On fait varier leur valeur entre deux bornes et on enregistre la valeur des variables résultat afin de savoir comment chaque variable aléatoire et chaque variable scénario influent sur la valeur des variables résultat. Pour une variable aléatoire, les valeurs explorées sont soit autour de la médiane, soit autour de la valeur par

défaut de la cellule, avec des limites définies par des percentiles ou des % de déviation. Pour une variable scénario, l'analyse est effectuée entre les deux bornes spécifiées lors de la définition de la variable. Le nombre de points étudiés pour les variables est une option qui peut être modifiée par l'utilisateur avant de lancer les simulations.

Le graphique araignée ne se limite pas à afficher le maximum et le minimum de changement de la variable résultat. Il représente la valeur de la variable résultat pour chacun des points étudiés des variables d'entrée. Cette approche est utile pour vérifier quel est le degré de dépendance entre les variables aléatoires ou scénario et les variables résultat et notamment vérifier si les relations sont monotones.

Options

Pour afficher la boîte de dialogue des options, cliquez sur le bouton  de la barre d'outils "XLSTAT-SIM". Utilisez cette boîte de dialogue pour définir les options générales du module XLSTAT-SIM.

Onglet **Générales** :

Modèle limité à : cette option permet de définir où se trouve les différents objets du modèle de simulation. Afin de gagner en vitesse de calcul, dans la mesure du possible, essayez de restreindre un modèle à une unique feuille de calcul. Les options suivantes sont disponibles :

- **Feuille** : seuls les objets de la feuille Excel active sont utilisés pour la simulation, les autres feuilles sont ignorées.
- **Classeur** : les objets intervenant dans le modèle de simulation sont recherchés dans l'ensemble du classeur actif. Cette option permet de répartir le modèle sur plusieurs feuilles Excel.

Méthode d'échantillonnage : **c** ette option permet de choisir la méthode de génération des échantillons. Deux possibilités sont proposées :

- **Classique** : les échantillons sont générés en utilisant des simulations Monte Carlo.
- **Hypercubes latins** : les échantillons sont générés à l'aide de la méthode des hypercubes latins. Cette méthode divise la distribution de la variable aléatoire en sections de même taille, puis génère à l'intérieur de chaque section des échantillons de même taille. Cela conduit à une convergence plus rapide de la simulation. Vous pouvez choisir le nombre de **sections**. La valeur par défaut est 500.

Mémoire pas à pas : entrez le nombre maximum d'itérations simulations qui seront stockées pour le mode « Faire un pas de simulation » pour le calcul des statistiques. Si la limite est atteinte, un glissement de la fenêtre de calcul est effectué : lors d'une nouvelle itération, la première itération de la fenêtre est oubliée et la nouvelle est stockée. La valeur par défaut est 500. Cette valeur peut être plus grande si nécessaire.

Nombre d'itérations par pas : entrez la valeur du nombre d'itérations qui sont effectuées au cours d'un pas. La valeur par défaut est 1.

Conditions d'arrêt : activez cette option pour définir quand la convergence est atteinte.

- **Critère** : Choisissez le critère à utiliser pour déterminer si la convergence est atteinte ou non :
- **Moyenne** : la convergence sera déterminée en analysant l'évolution de la moyenne des « variables résultats » choisies (voir ci-dessous).
- **Ecart-type** : la convergence sera déterminée en analysant l'évolution de l'écart-type des « variables résultats » choisies (voir ci-dessous).
- **Percentile** : la convergence sera déterminée en analysant l'évolution d'un percentile donné des « variables résultats » choisies (voir ci-dessous). Entrez la valeur de ce **percentile**. La valeur par défaut est 90%.
- **Fréquence de test** : entrez le nombre d'itérations à réaliser entre deux vérifications de la convergence. Valeur par défaut : 100.
- **Convergence** : entrez la valeur en % de l'évolution minimale des critères en-deçà de laquelle la convergence est atteinte. Valeur par défaut : 3%.
- **Intervalle de confiance (%)** : Entrez la taille en % de l'intervalle de confiance calculé autour du critère choisi. La borne supérieure de l'intervalle de confiance est comparée à la valeur de la convergence définie ci-dessus, pour déterminer si les calculs doivent se poursuivre ou non. Valeur par défaut : 95%.
- **Monitoring de la convergence** : choisissez quelles variables résultats doivent être suivies pour déterminer si il y a convergence ou non. Deux options sont proposées :
- **Toutes les variables résultat** : toutes les « variables résultats » impliquées dans le modèle de simulation sont utilisées pour déterminer la convergence.
- **Variables résultat activées** : seules les « variables résultat » pour lesquelles le champ ConvCheck a une valeur différente de 0 sont utilisées pour déterminer la convergence.

Onglet **Format** :

Utilisez ces options pour définir le format des différents objets des modèles de simulation utilisés dans les feuilles Excel :

- **Distributions** : vous pouvez définir la couleur de la police et la couleur du fond des cellules dans lesquelles sont enregistrées les définitions des variables aléatoire d'entrée et des distributions correspondantes.
- **Variables scénario** : vous pouvez définir la couleur de la police et la couleur du fond des cellules dans lesquelles sont enregistrées les variables scénario.
- **Variables résultat** : vous pouvez définir la couleur de la police et la couleur du fond des cellules dans lesquelles sont enregistrées les définitions des variables résultat.
- **Statistiques** : vous pouvez définir la couleur de la police et la couleur du fond des cellules dans lesquelles sont enregistrées les définitions des statistiques.

Niveau de filtre de l'affichage : sélectionnez le niveau de détail pour l'affichage du rapport de simulation. C'est à ce niveau qu'est contrôlé l'affichage des statistiques descriptives et des histogrammes des différents objets du modèle :

- **Tout** : les détails sont affichés pour tous les objets du modèle.
- **Activés** : les détails sont uniquement affichés pour les objets qui ont une valeur du paramètre Visible égale à 1.
- **Aucun** : les détails ne sont pas affichés.

Barre d'outils

XLSTAT-Sim est doté d'une barre d'outils spécifique. Elle peut être affichée en cliquant sur la barre d'outils XLSTAT-Sim de la barre d'outils « XLSTAT ».

Cliquez sur cette icône pour définir une distribution (voir Définir une distribution pour plus de détails).

Cliquez sur cette icône pour définir une variable scénario (voir Définir une variable de décision pour plus de détails).

Cliquez sur cette icône pour définir une variable résultat (voir Définir une variable résultat pour plus de détails).

Cliquez sur cette icône pour définir une statistique (voir Définir une statistique pour plus de détails).

Cliquez sur cette icône pour réinitialiser le modèle de simulation et faire une première itération.

Cliquez sur cette icône pour effectuer un pas de simulation. Le nombre de simulations effectuées à chaque pas peut être spécifié dans la boîte des Options.

Cliquez sur cette icône pour lancer les simulations et afficher le rapport de résultats.

Cliquez sur cette icône pour exporter un modèle de simulation. Toutes les fonctions XLSTAT-Sim sont alors transformées en commentaires. Les formules dans les cellules sont stockées sous forme de commentaires de cellules, et les formulas sont remplacées par la valeur par défaut de la fonction ou par la formule la reliant à d'autres cellules dans le cas de XLSTAT_SimRes.

 Cliquez sur cette icône pour importer un modèle de simulation. Toutes les fonctions XLSTAT-Sim sont alors extraites depuis les cellules en commentaire et exportées comme formules dans les cellules correspondantes.

 Cliquez sur cette icône pour afficher la boîte des options XLSTAT-Sim.

Exemple

Des exemples démontrant comment utiliser XLSTAT-Sim sont disponibles sur le Centre d'aide XLSTAT :

http://www.xlstat.com/demo-sim1f.htm

http://www.xlstat.com/demo-sim2f.htm

http://www.xlstat.com/demo-sim3f.htm

http://www.xlstat.com/demo-sim4f.htm

Bibliographie

Vose, D. (2008). Risk Analysis – A Quantitative Guide, Third Edition, John Wiley & Sons, New York.

Définir une distribution

Utilisez cet outil pour associer une loi de distribution à une cellule Excel correspondant à une variable (ou quantité) pour laquelle il existe une incertitude quant à sa vraie valeur. La distribution statistique permet de représenter l'incertitude.

Dans cette section :

Description

Les variables aléatoires auxquelles on associe une loi de distribution statistique sont l'un des deux éléments fondamentaux du modèle de simulation, l'autre étant les variables résultat. Pour une description plus détaillée des modèles de simulation, vous pouvez consulter l'introduction de XLSTAT-Sim.

Cet outil permet de définir la distribution théorique et les paramètres de cette loi qui seront utilisés pour générer un échantillon d'une variable aléatoire donnée. Un large choix de distribution est disponible (voir ci- dessous).

Pour définir la distribution suivie par une variable (physiquement, une cellule sur la feuille Excel), vous devez créer un appel à la fonction XLSTAT_SimX dialogue qui va générer pour vous la formule appelant XLSTAT_SimX. X correspond à la distribution choisie (voir le tableau si dessous).

Syntaxe de XLSTAT_SimX :

XLSTAT_SimX(VarName, Param1, Param2, Param3, Param4, Param5, TruncMode, LowerBound, UpperBound, DefaultType, DefaultValue, Visible)

VarName est une chaîne de caractères donnant le nom de la variable pour laquelle la distribution est définie ou une référence à la cellule contenant le nom de la variable. Le nom de la variable sera utilisé dans le rapport de simulation afin d'identifier la variable.

Param1 est un argument optionnel (la valeur par défaut est 0) qui donne la valeur du premier paramètre de la distribution choisie.

Param2 est un argument optionnel (la valeur par défaut est 0) qui donne la valeur du second paramètre de la distribution choisie.

Param3 est un argument optionnel (la valeur par défaut est 0) qui donne la valeur du troisième paramètre de la distribution choisie.

Param4 est un argument optionnel (la valeur par défaut est 0) qui donne la valeur du quatrième paramètre de la distribution choisie.

Param5 est un argument optionnel (la valeur par défaut est 0) qui donne la valeur du quatrième paramètre de la distribution choisie.

TruncMode est un entier optionnel qui indique si et comment la distribution est tronquée. Une valeur de 0 (valeur par défaut) correspond à ne pas tronquer la distribution. Une valeur de 1 correspond à tronquer la distribution entre deux bornes à entrer. Une valeur de 2 correspond à tronquer entre deux percentiles dont les valeurs doivent être entrées.

TruncLower est une valeur optionnelle indiquant la borne inférieure de la troncature.

TruncUpper est une valeur optionnelle indiquant la borne supérieure de la troncature.

DefaultType est un entier optionnel qui indique comment est déterminée la valeur par défaut de la variable : 0 (valeur par défaut) correspond à la moyenne théorique de la distribution étant donnés ses paramètres; 1 à la valeur donnée par l'argument DefaultValue.

DefaultValue est une valeur optionnelle prise en compte si DefaultType vaut 1, affichée dans la cellule comme résultat de la cellule lorsqu'aucun processus de simulation n'est en cours.

Visible est un argument optionnel qui indique si les détails concernant la variable aléatoire sont affichés dans le rapport de simulation dans le cas où le « Niveau d'affichage des résultats » est sur « Activés » dans la boîte des Options XLSTAT-Sim (onglet Format). 0 désactive l'affichage et 1 active l'affichage. La valeur par défaut est 1.

Distribution	Nom XLSTAT	Param1	Param2	Param3	Param4	Param5
Arcsinus	XLSTAT_SimArcsineG	alpha				
Bêta	XLSTAT_SimBeta	alpha	beta			
Bernoulli	XLSTAT_SimBernoulli	p				
Bêta	XLSTAT_SimBeta	alpha	beta			
Bêta4	XLSTAT_SimBeta4	alpha	beta	c	d	
Binomiale	XLSTAT_SimBinomial	n	p			
Khi²	XLSTAT_SimChiSqr	df				
Erlang	XLSTAT_SimErlang	k	gamma			
Exponentielle	XLSTAT_SimExponential	Lambda				
Fisher	XLSTAT_SimFisher	df1	df2			
Fisher-Tippett (1)	XLSTAT_SimFisherTippett1	beta				
Fisher-Tippett (2)	XLSTAT_SimFisherTippett2	beta	mu			
Gamma (1)	XLSTAT_SimGamma1	k				
Gamma (2)	XLSTAT_SimGamma2	k	beta			
Gamma (3)	XLSTAT_SimGamma3	k	beta	mu		
GEV	XLSTAT_SimGEV	beta	k	mu		
Gumbel	XLSTAT_SimGumbel					
Logistique	XLSTAT_SimLogistic	mu	sigma			
Lognormale	XLSTAT_SimLognormal	mu	sigma			
Lognormale2	XLSTAT_SimLognormal2	mu	s			
Binomiale négative (1)	XLSTAT_SimNegBinomial1	n	p			
Binomiale négative (2)	XLSTAT_SimNegBinomial2	k	p			
Normale	XLSTAT_SimNormal	mu	sigma			
Normale (Standard)	XLSTAT_SimNormalStd					
Pareto	XLSTAT_SimPareto	a	b			
Pert	XLSTAT_SimPert	a	m	b		
Poisson	XLSTAT_SimPoisson	Lambda				
Student	XLSTAT_SimStudent	df				
Trapezoïdale	XLSTAT_SimTrapezoidal	a	b	c	d	
Triangulaire	XLSTAT_SimTriangular	a	m	b		
TriangulaireQ	XLSTAT_SimTriangularQ	a	m	b	q1	q2
Uniforme	XLSTAT_SimUniform	a	b			
Uniforme discrète	XLSTAT_SimUniformDisc	a	b			
Weibull (1)	XLSTAT_SimWeibull1	beta				
Weibull (2)	XLSTAT_SimWeibull2	beta	gamma			
Weibull (3)	XLSTAT_SimWeibull3	beta	gamma	mu		

Exemple:

=XLSTAT_SimNormal("Revenue Q1", 50000, 5000)

Cette fonction associe une distribution normale avec une moyenne de 50000 et un écart-type de 5000 aux cellules où elles sont entrées. La cellule affiche 50000 (la valeur par défaut correspondant à la moyenne). Si un rapport est ensuite généré, les résultats correspondant à cette cellule seront identifiés par "Recettes Q1". Les Param3, Param4 et Param5 ne sont pas entrés parce que la distribution normale est définie par deux paramètres. Les autres paramètres ne sont pas entrés non plus et vaudront donc leur valeur par défaut.

Détermination des paramètres de la loi

En général, le choix de la loi et des paramètres de la loi est guidé par une connaissance empirique du phénomène, des résultats déjà disponibles ou encore des hypothèses de travail.

Pour choisir la loi la mieux adaptée et les paramètres qui conviennent que vous pouvez aussi utiliser l'outil d'ajustement à une loi de distribution de XLSTAT. Cet outil vous permet de calculer, sur la base d'un échantillon disponible, les meilleures valeurs des paramètres pour une loi donnée.

Distributions aléatoires disponibles dans XLSTAT-Sim

XLSTAT permet l'utilisation des lois suivantes :

- Arcsinus (a) : la densité de cette loi (dérivée de la loi Bêta de type I) est donnée par :

$$f(x) = \frac{\sin(\pi\alpha)}{\pi x}(\frac{x}{1-x})^{\alpha-1}, \quad \text{avec } 0 < \alpha < 1, x \in [0,1]$$

On a $E(X) = \alpha$ et $V(X) = \alpha(1-\alpha)/2$

- Bernoulli (p) : la densité de cette loi est donnée par :

$$P(X=1) = p, P(X=0) = 1-p, \quad \text{avec } p \in [0,1]$$

On a $E(X) = p$ et $V(X) = p(1-p)$

La loi de Bernoulli, du nom du mathématicien suisse Jacob Bernoulli (1654-1705), permet de décrire les phénomènes aléatoires binaires où seuls deux événements peuvent survenir avec des probabilités respectives de p et $1-p$.

- Bêta (a, b) : la densité de cette loi (aussi appelée Bêta de type I) est donnée par :

$$f(x) = \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}$$

$$\text{avec} \quad \alpha, \beta > 0, x \in [0,1] \text{ et } B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

On a $E(X) = \alpha/(\alpha+\beta)$ et $V(X) = \alpha\beta/[(\alpha+\beta+1)(\alpha+\beta)^2]$

- Bêta4 (α, β, c, d) : la densité de cette loi est donnée par :

$$f(x) = \frac{1}{B(\alpha,\beta)} \frac{(x-c)^{\alpha-1}(d-x)^{\beta-1}}{(d-c)^{\alpha+\beta-1}}, \quad \text{avec } \alpha, \beta > 0, x \in [c,d]$$

$$c, d \in \mathbb{R} \text{ et } B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

On a $E(X) = \frac{c+(c-d)\alpha}{(\alpha+\beta)}$ et $V(X) = \frac{(c-d)^2\alpha\beta}{(\alpha+\beta+1)(\alpha+\beta)^2}$

Pour la loi Bêta de type I, la distribution est dans l'intervalle $[0, 1]$. La loi Bêta4 est obtenue par un simple changement de variable de la loi Bêta de type I de telle sorte que la distribution soit sur l'intervalle [c, d].

- Binomiale (n, p) : la densité de cette loi est donnée par :

$$P(X=x) = C_n^x p^x (1-p)^{n-x}, \quad \text{avec } n, x \in \mathbb{N}, n > 0, x \in [0,n], p \in [0,1]$$

On a $E(X) = np$ et $V(X) = np(1-p)$

n est le nombre d'essais, et p la probabilité de succès. La loi binomiale est la loi du nombre de succès pour n essais, sachant que la probabilité de succès vaut p. La loi binomiale peut être vue comme la loi de n tirages dans une loi de Bernoulli.

- Binomiale négative (n, p) de type I : la densité de cette loi est donnée par :

$$P(X = x) = C_{n+x-1}^{x-1} p^n (1-p)^x, \quad \text{avec } n, x \in \mathbb{N}, n > 0, p \in [0, 1]$$

On a $E(X) = n(1-p)/p$ et $V(X) = n(1-p)/p^2$

n est le nombre de succès et p la probabilité de succès. La loi binomiale négative de type I est la loi du nombre de tirages x sans succès nécessaires avant d'avoir obtenus n succès.

- Binomiale négative (k, p) de type II : la densité de cette loi est donnée par :

$$P(X = x) = \frac{\Gamma(k+x) p^x}{x! \Gamma(k) (1+p)^{k+x}}, \quad \text{avec } x \in N, k, p > 0$$

On a $E(X) = kp$ et $V(X) = kp(p+1)$

La loi binomiale négative de type II permet de représenter des phénomènes discrets fortement hétérogènes. Lorsque k tend vers l'infini, la loi binomiale négative de type II tend vers une loi de Poisson de paramètre $\lambda = kp$.

- $Khi^2(df)$: la densité de cette loi est donnée par :

$$f(x) = \frac{(1/2)^{df/2}}{\Gamma(df/2)} x^{\frac{df}{2}-1} e^{-x/2}, \quad \text{avec } x > 0, df \in \mathbb{N}^*$$

On a $E(X) = df$ et $V(X) = 2df$

La loi du Khi² correspond à la loi de la somme des carrés de df lois normales centrées réduites (lois normales standard). Elle est très utilisée pour tester des hypothèses.

- Erlang (k, λ) : la densité de cette loi est donnée par :

$$f(x) = \lambda^k x^{k-1} \frac{e^{-\lambda x}}{(k-1)!}, \quad \text{avec } x \geq 0 \text{ et } k, \lambda > 0 \text{ et } k \in \mathbb{N}$$

On a $E(X) = k/\lambda$ et $V(X) = k/\lambda^2$

k est le paramètre de forme de la loi et λ est le paramètre de taux.

Cette distribution, développée par le scientifique danois A. K. Erlang (1878-1929) pour l'étude du trafic téléphonique, est utilisée de manière plus générale pour l'étude des files d'attente.

Remarque : lorsque $k = 1$, cette distribution est équivalente à la distribution exponentielle, et la loi Gamma à deux paramètres est une généralisation de la loi d'Erlang au cas où k est un réel et non un entier (par ailleurs on utilise le paramètre d'échelle $\beta = 1/\lambda$).

- Exponentielle (λ) : la densité de cette loi est donnée par :

$$f(x) = \lambda \exp(-\lambda x), \quad \text{avec } x > 0 \text{ et } \lambda > 0$$

On a $E(X) = 1/\lambda$ et $V(X) = 1/\lambda^2$

La loi exponentielle est souvent utilisée pour étudier la durée de vie en contrôle qualité.

- Fisher (df_1, df_2) : la densité de cette loi est donnée par :

$$f(x) = \frac{1}{xB(df_1/2, df_2/2)} \left(\frac{df_1 x}{df_1 x + df_2}\right)^{df_1/2} \left(1 - \frac{df_1 x}{df_1 x + df_2}\right)^{df_2/2}$$

$$\text{avec } x > 0 \text{ et } df_1, df_2 \in \mathbb{N}^*$$

On a $E(X) = df_2/(df_2 - 2)$ si $df_2 > 2$, et $V(X) = \frac{2df_2^2(df_1 + df_2 - 2)}{df_1(df_2-2)^2(df_2-4)}$ si $df_2 > 4$

La loi de Fisher, du nom du biologiste, généticien et statisticien Ronald Aylmer Fisher (1890-1962), correspond au rapport de deux lois du Khi². Elle est très utilisée pour tester des hypothèses.

- Fisher-Tippett (β, μ) : la densité de cette loi est donnée par :

$$f(x) = \frac{1}{\beta} \exp\left(-\frac{x-\mu}{\beta} - \exp\left(-\frac{x-\mu}{\beta}\right)\right), \quad \text{avec } \beta > 0$$

On a $E(X) = \mu + \beta\gamma$ et $V(X) = (\pi\beta)^2/6$ où γ est la constante de Euler-Mascheroni.

La loi de Fisher-Tippett, aussi appelée loi Log-Weibull, ou loi généralisée des valeurs extrêmes, est utilisée dans l'étude de phénomènes extrêmes. La loi de Gumbel est un cas particulier de la loi de Fisher-Tippett avec $\beta = 1$ et $\mu = 0$.

- Gamma (k, β, μ) : la densité de cette loi est donnée par :

$$f(x) = (x-\mu)^{k-1} \frac{e^{-(x-\mu)/\beta}}{\beta^k \Gamma(k)}, \quad \text{avec } x > \mu \text{ et } k, \beta > 0$$

On a $E(X) = \mu + k\beta$ et $V(X) = k\beta^2$

k est le paramètre de forme de la loi et β est le paramètre d'échelle.

- GEV (β, k, μ): la densité de cette loi est donnée par :

$$f(x) = \frac{1}{\beta}\left(1 + k\frac{x-\mu}{\beta}\right)^{-1/k-1} \exp\left(-\left(1 + k\frac{x-\mu}{\beta}\right)^{-1/k}\right), \quad \text{avec} \quad \beta > 0$$

On a $E(X) = \mu + \frac{\beta}{k}\Gamma(1+k)$ et $V(X) = \left(\frac{\beta}{k}\right)^2 \left(\Gamma(1+2k) - \Gamma^2(1+k)\right)$

La loi GEV (Generalized Extreme Values) est très utilisée en hydrologie pour modéliser les phénomènes de crues. k est classiquement compris entre -0.6 et 0.6.

- Gumbel : la densité de cette loi est donnée par :

$$f(x) = \exp(-x - \exp(-x))$$

On a $E(X) = \gamma$ et $V(X) = \pi^2/6$ où γ est la constante de Euler-Mascheroni (0.5772156649…).

La loi de Gumbel, du nom de Emil Julius Gumbel (1891-1966), est un cas particulier de la loi de Fisher-Tippett avec $\beta = 1$ et $\mu = 0$. Elle est utilisée dans l'étude de phénomènes extrêmes comme les précipitations ou les crues maximales et les magnitudes maximales de tremblement de terre.

- Logistique (μ, s) : la densité de cette loi est donnée par :

$$f(x) = \frac{e^{-\frac{(x-\mu)}{s}}}{s(1 + e^{-\frac{(x-\mu)}{s}})}, \quad \text{avec } s > 0$$

On a $E(X) = \mu$ et $V(X) = (\pi s)^2/3$

- Lognormale (μ, σ) : la densité de cette loi est donnée par :

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(\ln(x)-\mu)^2}{2\sigma^2}}, \quad \text{avec } x, \sigma > 0$$

On a $E(X) = exp(\mu + \sigma^2/2)$ et $V(X) = \left(exp(\sigma^2) - 1\right) exp(2\mu + \sigma^2)$

- Lognormale2 (m, s) : la densité de cette loi est donnée par :

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(\ln(x)-\mu)^2}{2\sigma^2}}, \text{ avec } x, \sigma > 0$$

On a :

$\mu = ln(m) - ln(1 + s^2/m^2)/2$ et $\sigma^2 = ln(1 + s^2/m^2)$

Et :

$E(X) = m$ et $V(X) = s^2$

Cette distribution est simplement une reparamétrisation de la loi Lognormale.

- Normale (μ, σ) : la densité de cette loi est donnée par :

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad \text{avec } \sigma > 0$$

On a $E(X) = \mu$ et $V(X) = \sigma^2$

- Normale standard : la densité de cette loi est donnée par :

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

On a $E(X) = 0$ et $V(X) = 1$

Cette loi est un cas particulier de la loi normale, avec $\mu = 0$ et $\sigma = 1$. Elle est aussi appelée loi normale centrée réduite.

- Pareto (a, b) : la densité de cette loi est donnée par :

$$f(x) = \frac{ab^a}{x^{a+1}}, \text{ avec } a, b > 0 \text{ et } x \geq b$$

On a $E(X) = ab/(a-1)$ et $V(X) = \frac{ab^2}{((a-1)^2(a-2))}$

La loi de Pareto, du nom de l'économiste italien Vilfredo Pareto (1848-1923), est aussi connue sous le nom de loi de Bradford. Cette loi a d'abord été utilisée pour représenter la répartition des richesses dans la société, avec notamment le principe de Pareto, selon lequel 80% des richesses d'un pays sont détenus par 20% de la population.

- PERT (a, m, b) : la densité de cette loi est donnée par :

$$f(x) = \frac{1}{B(\alpha, \beta)} \frac{(x-a)^{\alpha-1}(b-x)^{\beta-1}}{(b-a)^{\alpha+\beta-1}}, \text{ avec } \alpha, \beta > 0, x \in [a, b]$$

$$a, b \in \mathbb{R} \text{ et } B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

$$\alpha = \frac{4m + b - 5a}{b - a}$$

$$\beta = \frac{5b - a - 4m}{b - a}$$

On a $E(X) = (b-a)\alpha/(\alpha+\beta)$ et $V(X) = (b-a)^2\alpha\beta/\left((\alpha+\beta+1)\left(\alpha+\beta\right)^2\right)$

La loi de PERT est donc un cas particulier de la loi Bêta4, définie par son intervalle de définition $[a, b]$ et sa valeur la plus probable m (le mode). PERT est l'acronyme de *Program Evaluation*

and Review Technique , une méthode de gestion et de planification de projet. La méthodologie et la distribution PERT ont été utilisées pour la première fois pour le projet de développement des missiles Polaris lancés depuis des sous-marins par la marine américaine et Lockheed de 1956 à 1960 (Clark 1962). La distribution PERT permet de modéliser le temps probable nécessaire à une équipe pour terminer son projet. La loi triangulaire, plus simple, permet aussi de modéliser ce type de phénomènes avec les trois mêmes paramètres.

- Poisson (λ): la densité de cette loi est donnée par :

$$P(X = x) = \frac{\exp(-\lambda)\lambda^x}{x!}, \quad \text{avec } x \in \mathbb{N} \text{ et } \lambda > 0$$

On a $E(X) = \lambda$ et $V(X) = \lambda$

La loi de Poisson, découverte par le mathématicien et astronome Siméon-Denis Poisson (1781-1840) qui fut élève de Laplace, Lagrange et Legendre, est souvent utilisée pour étudier des phénomènes de file d'attente.

- Student (df) : la densité de cette loi est donnée par :

$$f(x) = \frac{\Gamma((df + 1/2))}{\sqrt{\pi df}\Gamma(df/2)}(1 + x^2/df)^{-(df+1)/2}, \text{ avec } df > 0$$

On a $E(X) = 0$ si $df > 1$ et $V(X) = df/(df - 2)$ si $df > 2$

La loi de Student, du nom que se donnait le chimiste et statisticien anglais William Sealy Gosset (1876-1937) afin de préserver son anonymat (la brasserie Guinness interdisait à ses employés de publier, suite à la publication par un autre chercheur d'informations confidentielles) est la loi de la moyenne de df variables distribuées suivant une loi normale centrée réduite. Lorsque $df = 1$, la loi de Student est une loi de Cauchy dont la particularité est de n'avoir ni espérance ni variance.

- Trapézoïdale (a, b, c, d) : la densité de cette loi est donnée par :

$$\begin{cases} f(x) = \frac{2(x-a)}{(d+c-b-a)(b-a)}, & x \in [a, b] \\ \\ f(x) = \frac{2}{(d+c-b-a)}, & x \in [b, c] \\ \\ f(x) = \frac{2(d-x)}{(d+c-b-a)(d-c)}, & x \in [c, d] \\ \\ f(x) = 0, & x < a, \quad x > d \\ \\ \text{avec } a < b < c < d \end{cases}$$

On a $E(X) = \frac{d^2+c^2-b^2-a^2+cd-ab}{3(d+c-b-a)}$ et $V(X) = \frac{(c+d)(c^2+d^2)-(a+b)(a^2+b^2)}{6(d+c-b-a)} - E^2(X)$

Cette loi est utile pour représenter un phénomène dont on sait qu'il peut prendre des valeurs entre deux extrêmes, mais pour lequel un intervalle plus restreint paraît plus raisonnable.

- Triangulaire (a, m, b) : la densité de cette loi est donnée par :

$$\begin{cases} f(x) = \frac{2(x-a)}{(b-a)(m-a)}, & x \in [a, m] \\ f(x) = \frac{2(b-x)}{(b-a)(b-m)}, & x \in [m, b] \\ f(x) = 0, & x < a, \quad x < b \\ \text{avec } a < m < b \end{cases}$$

On a $E(X) = (a + m + b)/3$ et $V(X) = (a^2 + m^2 + b^2 - ab - am - bm)/18$

- TriangulaireQ (q_1, m, q_2, p_1, p_2): cette loi est une reparamétrisation de la loi triangulaire. Une première étape nécessite l'estimation des paramètres a et b de la distribution triangulaire pour savoir à quels quantiles q_1 et q_2 correspondent les pourcentages p_1 et p_2. Une fois ceci fait, on peut utiliser la fonction de densité ou de répartition triangulaire.

- Uniforme (a, b) : la densité de cette loi est donnée par :

$$f(x) = \frac{1}{b-a}, \text{ avec } b > a \text{ et } x \in [a, b]$$

On a $E(X) = (a + b)/2$ et $V(X) = (b - a)^2/12$

La loi uniforme (0, 1) est très utilisée pour les simulations. Comme la fonction de répartition de toutes les lois est comprise entre 0 et 1, un échantillon tiré dans une loi Uniforme (0,1) permet d'obtenir un échantillon dans toutes les lois dont on sait calculer l'inverse.

- Uniforme discrète (a, b) : la densité de cette loi est donnée par :

$$P[X = x] = \frac{1}{b-a+1}, \text{ avec } (a, b, x) \in \mathbb{N}^3, x \in [a, b]$$

On a $E(X) = (a + b)/2$ et $V(X) = [(b - a + 1)^2 - 1]/12$

La loi uniforme discrète correspond au cas particulier où la loi uniforme est restreinte à des nombre entiers.

- Weibull (β) : la densité de cette loi est donnée par :

$$f(x) = \beta x^{\beta-1} \exp(-x^\beta), \text{ avec } x > 0 \text{ et } \beta > 0$$

On a $E(X) = \Gamma(\frac{1}{\beta} + 1)$ et $V(X) = \Gamma(\frac{2}{\beta} + 1) - \Gamma^2(\frac{1}{\beta} + 1)$

Le paramètre β est le paramètre de forme de la loi de Weibull.

- Weibull (β, γ) : la densité de cette loi est donnée par :

$$f(x) = \frac{\beta}{\gamma}(\frac{x}{\gamma})^{\beta-1} e^{-(\frac{x}{\gamma})^{\beta}}, \text{ avec } x > 0, \text{ et } \beta, \gamma > 0$$

On a $E(X) = \gamma\Gamma\left(\frac{1}{\beta}+1\right)$ et $V(X) = \gamma^2\left[\Gamma\left(\frac{2}{\beta}+1\right) - \Gamma^2\left(\frac{1}{\beta}+1\right)\right]$

Le paramètre β est le paramètre de forme et le paramètre γ est le paramètre d'échelle. Lorsque $\beta = 1$, la loi de Weibull est une loi exponentielle de paramètre $1/\gamma$.

- Weibull (β, γ, μ): la densité de cette loi est donnée par :

$$f(x) = \frac{\beta}{\gamma}\left(\frac{x-\mu}{\gamma}\right)^{\beta-1} e^{-(\frac{x-\mu}{\gamma})^{\beta}}, \text{ avec } x > \mu, \text{ et } \beta, \gamma > 0$$

On a $E(X) = \mu + \gamma\Gamma\left(\frac{1}{\beta}+1\right)$ et $V(X) = \gamma^2\left[\Gamma\left(\frac{2}{\beta}+1\right) - \Gamma^2\left(\frac{1}{\beta}+1\right)\right]$

La loi de Weibull, du nom du suédois Ernst Hjalmar Waloddi Weibull (1887-1979), est très utilisée en contrôle qualité et en analyse de survie. Le paramètre β est le paramètre de forme et le paramètre γ est le paramètre d'échelle. Lorsque $\beta = 1$ et $\mu = 0$, la loi de Weibull est une loi exponentielle de paramètre $1/\gamma$.

Boîte de dialogue

OK : cliquez sur ce bouton pour lancer les calculs.

Annuler : cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

Aide : cliquez sur ce bouton pour afficher l'aide.

 : cliquez sur ce bouton pour rétablir les options par défaut.

 : cliquez sur ce bouton pour effacer les sélections de données.

Onglet **Général**:

Nom de la variable : entrez le nom de la variable aléatoire ou sélectionnez une cellule où se trouve le nom de la variable. Si vous sélectionnez une cellule, XLSTAT crée une référence absolue (par exemple, $A4$) ou une référence relative (par exemple, A4) en fonction de votre choix au niveau des options générales de XLSTAT.

Distributions : choisissez la distribution que vous voulez affecter à la variable aléatoire (voir la section description pour plus d'information sur les distributions disponibles).

Paramètres : entrez la valeur des paramètres de la distribution que vous avez choisie.

Troncature : activez cette option pour tronquer la distribution.

- **Absolue** : choisissez cette option si vous voulez entrer les bornes de la troncature sous forme de valeurs de la variable (par exemple -2 et 2 pour une loi normale standard).
- **Percentile** : choisissez cette option si vous voulez entrer les bornes de la troncature sous forme de percentiles (par exemple 10% et 90%).
- **Borne inférieure** : Entrez la valeur de la borne inférieure de l'intervalle de troncature.
- **Borne supérieure** : Entrez la valeur de la borne inférieure de l'intervalle de troncature.

Onglet **Options**:

Valeur par de défaut de la cellule : choisissez la valeur par défaut de la variable aléatoire. Cette valeur sera retournée comme valeur de la cellule lorsqu'une simulation n'est pas en cours. La valeur peut être définie par l'une des trois méthodes suivantes :

- **Espérance** : l'espérance mathématique (ou moyenne théorique) de la distribution est retournée.
- **Valeur fixe** : entrez la valeur par défaut de la variable aléatoire.
- **Référence** : sélectionnez une cellule active dans la feuille Excel qui contient la valeur par défaut.

Afficher les résultats : activez cette option pour afficher les résultats détaillés concernant la variable dans le rapport de simulation. Cette option n'est active que si vous avez sélectionné le niveau limité de filtre d'affichage « Activés » dans la boîte d'Options de XLSTAT-Sim (voir la section Options pour plus de détails).

Résultats

Un appel à la fonction XLSTAT_SimX est généré dans la cellule active, avec les paramètres entrés. La formule suivante remplace alors le contenu de la cellule :

= XLSTAT_SimX(VarName, Param1, Param2, Param3, Param4, Param5, TruncMode, LowerBound, UpperBound, DefaultType, DefaultValue, Visible)

La couleur du fond et la couleur de la police de la cellule sont celles définies dans les options de XLSTAT-Sim.

Définir une variable scénario

Utilisez cet outil pour définir une variable dont la valeur varie entre deux limites connues au cours de l'analyse tornado.

Dans cette section :

Description

Cet outil permet de construire une variable de scénario qui pourra ensuite être utilisée par XLSTAT-Sim pour l'analyse tornado. Pour plus de détails sur la construction d'un modèle de simulation, référez-vous à la section introduction de XLSTAT-Sim.

Une variable scénario est utilisée pour l'analyse tornado. Cette fonction vous donne la possibilité de définir une variable scénario en renseignant les bornes entre lesquelles elle doit varier.

Pour définir la variable scénario (physiquement, une cellule sur la feuille Excel), vous devez créer un appel à la fonction XLSTAT_SimSVar ou utiliser la boîte de dialogue qui va générer pour vous la formule appelant XLSTAT_SimSVar.

Syntaxe de XLSTAT_SimSVar

XLSTAT_SimSVar (SVarName, LowerBound, UpperBound, Mode, Step, DefaultType, DefaultValue, Visible)

SVarName est une chaîne de caractères donnant le nom de la variable scénario. Le nom de la variable sera utilisé dans le rapport de simulation afin d'identifier la variable.

LowerBound correspond à la borne inférieure de l'intervalle de la variable scénario.

UpperBound correspond à la borne supérieure de l'intervalle de la variable scénario.

Type est un entier qui indique le type de données de la variable scenario. 1 correspond à une variable continue, et 2 à une variable discrète. La valeur par défaut est 1.

Step est un nombre qui indique dans le cas d'une variable discrète quel doit être l'écart entre deux points de la variable scénario pour l'analyse tornado. Ce paramètre est optionnel avec une valeur par défaut de 1.

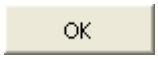
DefaultType est un entier optionnel qui indique comment est déterminée la valeur par défaut de la variable : 0 (valeur par défaut) correspond à la moyenne théorique de la distribution étant

donnés ses paramètres; 1 à la valeur donnée par l'argument DefaultValue.

DefaultValue est un nombre donnant la valeur par défaut de la variable scénario. Cette valeur est utilisée comme la valeur de la variable excepté pendant l'analyse tornado. Elle est toujours affichée comme valeur de la cellule qui contient la fonction.

Visible est un argument optionnel qui indique si les détails concernant la variable scénario sont affichés dans le rapport de simulation dans le cas où le « Niveau d'affichage des résultats » est sur « Activés » dans la boîte des Options XLSTAT-Sim (onglet Format). 0 désactive l'affichage et 1 active l'affichage. La valeur par défaut est 1.

Boîte de dialogue



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.

Onglet **Général**:

Nom de la variable : entrez le nom de la variable scénario ou sélectionnez une cellule où se trouve le nom de la variable. Si vous sélectionnez une cellule, XLSTAT crée une référence absolue (par exemple, $A4$) ou une référence relative (par exemple, A4) en fonction de votre choix au niveau des options générales de XLSTAT.

Valeur par défaut : choisissez la valeur par défaut de la variable scénario. Cette valeur sera retournée comme valeur de la cellule lorsqu'une simulation n'est pas en cours. La valeur peut être définie par l'une des deux méthodes suivantes :

- **Référence** : sélectionnez une cellule active dans la feuille Excel qui contient la valeur par défaut.
- **Valeur fixe** : entrez la valeur par défaut de la variable scénario.

Borne inférieure : entrez la valeur de la borne inférieure ou sélectionnez une cellule qui contient la valeur de la borne inférieure.

Borne supérieure : entrez la valeur de la borne supérieure ou sélectionnez une cellule qui contient la valeur de la borne supérieure.

Type de données :

Continues : Choisissez cette option pour définir une variable scénario continue pouvant prendre toute valeur dans l'intervalle compris entre les bornes supérieures et inférieures.

Discrètes : Choisissez cette option pour définir une variable scénario discrète.

- **Pas** : entrez la valeur du pas ou sélectionnez une cellule qui contient la valeur du pas.

Onglet **Options**:

Valeur par de défaut de la cellule : choisissez la valeur par défaut de la variable scénario. Cette valeur sera retournée comme valeur de la cellule lorsqu'une simulation n'est pas en cours. La valeur peut être définie par l'une des trois méthodes suivantes :

- **Espérance** : l'espérance mathématique (ou moyenne théorique) de la distribution est retournée.
- **Valeur fixe** : entrez la valeur par défaut de la variable aléatoire.
- **Référence** : sélectionnez une cellule active dans la feuille Excel qui contient la valeur par défaut.

Afficher les résultats : activez cette option pour afficher les résultats détaillés concernant la variable dans le rapport de simulation. Cette option n'est active que si vous avez sélectionné le niveau limité de filtre d'affichage « Activés » dans la boîte d'Options de XLSTAT-Sim (voir la section Options pour plus de détails).

Résultats

Un appel à la fonction XLSTAT_SimSVar est généré dans la cellule active, avec les paramètres entrés. La formule suivante remplace alors le contenu de la cellule :

=XLSTAT_SimSVar(SVarName, LowerBound, UpperBound, Type, Step, DefaultType, DefaultValue, Visible)

La couleur du fond et la couleur de la police de la cellule sont celles définies dans les options de XLSTAT-Sim.

Définir une variable résultat

Utilisez cet outil pour définir une variable résultat dont le calcul est justement le but même du modèle de simulation.

Dans cette section :

Description

Les variables résultat sont avec les variables aléatoires auxquelles on associe une loi de distribution l'un des deux éléments fondamentaux du modèle de simulation. Pour une description plus détaillée des modèles de simulation, vous pouvez consulter l'introduction de XLSTAT-Sim.

Si vous le souhaitez, les variables résultat peuvent être utilisées pour définir le moment où il faut interrompre les simulations : si, dans la boîte de dialogue des Options XLSTAT-Sim, vous avez demandé que les "Variables résultat actives" soient utilisées pour arrêter les simulations, par exemple lorsque la moyenne a convergée, alors, si le paramètre ConvActiv de la variable résultat est fixé à 1, la moyenne de la variable sera utilisée pour déterminer si les simulations ont convergé ou non.

Pour définir la variable résultat (physiquement, une cellule sur la feuille Excel), vous devez créer un appel à la fonction XLSTAT_SimRes ou utiliser la boîte de dialogue qui va générer pour vous la formule appelant XLSTAT_SimRes.

Syntaxe de XLSTAT_SimRes :

XLSTAT_SimRes (ResName, Formula, DefaultValue, ConvActiv, Visible)

ResName est une chaîne de caractères donnant le nom de la variable résultat ou une référence à la cellule contenant le nom de la variable. Le nom de la variable sera utilisé dans le rapport de simulation afin d'identifier la variable.

Formule est une chaîne contenant la formule qui sera utilisée pour le calcul de la variable résultat. La formule permet de relier directement ou in directement la variable résultat aux variables aléatoires et éventuellement aux variables scénario. La formule doit être une formule Excel non précédée du "=".

DefaultValue est un nombre donnant la valeur par défaut de la variable résultat. Cette valeur n'est pas utilisée et est seulement affichée par défaut dans la cellule contenant la formule.

ConvActiv est un entier qui indique si cette variable résultat est analysée lors des tests de convergence. Cette option n'est active que si l'option de convergence « Variables résultat activées » est sur « Variables résultat activées » (voir les Options XLSTAT-Sim). La valeur doit être 1 si vous souhaitez que la variable soit utilisée pour le monitoring des résultats. La valeur par défaut est 0.

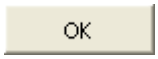
Visible est un argument optionnel qui indique si les détails concernant la variable scénario sont affichés dans le rapport dans le cas où « Niveau d'affichage des résultats » est sur « Activés » dans la boîte des Options XLSTAT-Sim (onglet Format). 0 désactive l'affichage et 1 active l'affichage. La valeur par défaut est 1.

Exemple:

=XLSTAT_SimRes("Prévision N+1", B3+B4-B5)

Cette fonction définit dans la cellule active une variable résultat nommée « Prévision N+1 » calculée comme la somme des cellules B3+B4 moins la cellule B5 et ayant une valeur par défaut égale à cette somme. Les autres paramètres ne sont pas entrés non plus et vaudront donc leur valeur par défaut.

Boîte de dialogue



: cliquez sur ce bouton pour créer la variable.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer de modification.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.

Onglet **Général**:

Nom de la variable : entrez le nom de la variable résultat ou sélectionnez une cellule où se trouve le nom de la variable. Si vous sélectionnez une cellule, XLSTAT crée une référence absolue (par exemple, $A4$) ou une référence relative (par exemple, A4) en fonction de votre choix au niveau des options générales de XLSTAT.

Monitoring de la convergence : Activez cette option pour inclure cette variable résultat aux variables utilisées pour tester la convergence. Cette option n'est active que si vous avez sélectionné «Variables résultat activées » dans les options de convergence du menu XLSTAT-Sim.

Afficher les résultats : activez cette option pour afficher les résultats détaillés concernant la variable dans le rapport de simulation. Cette option n'est active que si vous avez sélectionné le niveau limité de filtre d'affichage « Activés » dans la boîte d'Options de XLSTAT-Sim (voir la section Options pour plus de détails).

Résultats

Un appel à la fonction XLSTAT_SimRes est généré dans la cellule active, avec les paramètres entrés. La formule suivante remplace alors le contenu de la cellule :

XLSTAT_SimRes (ResName, DefaultValue, Formula, ConvActiv, Visible)

La couleur du fond et la couleur de la police de la cellule sont celles définies dans les options de XLSTAT-Sim.

Définir une statistique

Utilisez cet outil dans un modèle de simulation pour définir une statistique calculée pour une variable du modèle de simulation. La statistique est mise à jour après chaque itération du processus de simulation. Les résultats relatifs aux statistiques définies sont disponibles dans le rapport de simulation. Un large choix de statistiques est proposé.

Dans cette section :

Description

Cet outil permet de créer une variable qui est en fait une statistique calculée après chaque itération du processus de simulation. La statistique peut être calculée sur une variable aléatoire d'entrée ou sur une variable résultat.

Pour définir la statistique (physiquement, une cellule sur la feuille Excel), vous devez créer un appel à la fonction XLSTAT_SimStatX/TheoX/SPCX ou utiliser la boîte de dialogue qui va générer pour vous la formule appelant la fonction. X correspond à la statistique choisie. Les statistiques disponibles sont listées dans les tableaux ci-dessous.

Syntaxe de XLSTAT_SimStatX/TheoX/SPCX

XLSTAT_SimStatX(StatName, Reference, Visible)

XLSTAT_SimTheoX(StatName, Reference, Visible)

XLSTAT_SimSPCX(StatName, Reference, Visible)

X correspond à la statistique que vous voulez calculer, telle que définie dans les tableaux ci-dessous.

StatName est une chaîne de caractères donnant le nom de la statistique ou une référence à la cellule contenant le nom de la statistique. Le nom de la statistique sera utilisé dans le rapport de simulation afin d'identifier la variable.

Reference indique la variable sur laquelle doit être calculée la statistique. La référence doit correspondre à la cellule Excel contenant la variable en question.

Visible est un argument optionnel qui indique si les détails concernant la statistique sont affichés dans le rapport de simulation dans le cas où le « Niveau d'affichage des résultats » est

sur « Activés » dans la boîte des Options XLSTAT-Sim (onglet Format). 0 désactive l'affichage et 1 active l'affichage. La valeur par défaut est 1.

Statistiques descriptives

Les statistiques descriptives suivantes sont disponibles :

Statistique	Nom XLSTAT
Nombre d'observations	XLSTAT_SimStatNbrObs
Nombre de valeurs manquantes	XLSTAT_SimStatNbrMiss
Somme des poids	XLSTAT_SimStatSumOfWeights
Minimum	XLSTAT_SimStatMinimum
Maximum	XLSTAT_SimStatMaximum
Fréquence du minimum	XLSTAT_SimStatFreqMin
Fréquence du maximum	XLSTAT_SimStatFreqMax
Amplitude	XLSTAT_SimStatAmplitude
1ert quartile	XLSTAT_SimStat1stQuartile
Médiane	XLSTAT_SimStatMedian
3ème quartile	XLSTAT_SimStat3rdQuartile
Somme	XLSTAT_SimStatSum
Moyenne	XLSTAT_SimStatMean
Variance n	XLSTAT_SimStatVarianceN
Variance n-1	XLSTAT_SimStatVarianceN1
Ecart-type n	XLSTAT_SimStatStdevN
Ecart-type n-1	XLSTAT_SimStatStdevN1
Coefficient de variation	XLSTAT_SimStatVariation
Asymétrie (Pearson)	XLSTAT_SimStatSkewnessPearson
Asymétrie (Fisher)	XLSTAT_SimStatSkewnessFisher
Asymétrie (Bowley)	XLSTAT_SimStatSkewnessBowley
Aplatissement (Pearson)	XLSTAT_SimStatKurtosisPearson
Aplatissement (Fisher)	XLSTAT_SimStatKurtosisFisher
Ecart-type de la moyenne	XLSTAT_SimStatStdErrorOfMean
Borne inférieure de la moyenne	XLSTAT_SimStatLowerBound95Perc
Borne supérieure de la moyenne	XLSTAT_SimStatUpperBound95Perc
Ecart-type Asymétrie	XLSTAT_SimStatStdErrorOfSkewness
Ecart-type Aplatissement	XLSTAT_SimStatStdErrorOfKurtosis
Ecart absolu moyen	XLSTAT_SimStatMeanAbsDeviation
Ecart absolu médian	XLSTAT_SimStatMedianAbsDeviation
Moyenne géométrique	XLSTAT_SimStatGeometricMean
Ecart-type géométrique	XLSTAT_SimStatGSD
Moyenne harmonique	XLSTAT_SimStatHarmonicMean

Les détails et les formules relatives aux statistiques ci-dessus sont disponibles dans la section description de l'outil « Statistiques descriptives » de XLSTAT.

Statistiques théoriques

Ces statistiques sont calculées à partir des formules reliant à ces statistiques aux paramètres des lois.

Statistique	Nom XLSTAT
Moyenne	XLSTAT_SimTheoMean
Variance	XLSTAT_SimTheoVariance
Ecart-type	XLSTAT_SimTheoStdev

Autres statistiques

D'autres statistiques provenant du domaine de la maîtrise statistique des procédés (en anglais SPC pour Statistical Process Control) sont disponibles. Ces statistiques ne sont disponibles que pour les utilisateurs disposant d'une licence pour le module XLSTAT-SPC.

Statistique	Nom XLSTAT
Cp	XLSTAT_SimSPCCp
Cp lower	XLSTAT_SimSPCCpl
Cp upper	XLSTAT_SimSPCCpu
Cpk	XLSTAT_SimSPCCpk
Pp	XLSTAT_SimSPCPp
Pp lower	XLSTAT_SimSPCPpl
Pp upper	XLSTAT_SimSPCPpu
Ppk	XLSTAT_SimSPCPpk
Cpm	XLSTAT_SimSPCCpm
Cpm (Boyles)	XLSTAT_SimSPCCpmBoyles
Cp 5.5	XLSTAT_SimSPCCp55
Cpk 5.5	XLSTAT_SimSPCCpk55
Cpmk	XLSTAT_SimSPCCpmk
Cs (Wright)	XLSTAT_SimSPCCsWright
Z below	XLSTAT_SimSPCZbelow
Z above	XLSTAT_SimSPCZabove
Z total	XLSTAT_SimSPCZtotal
p(not conform) below	XLSTAT_SimSPCpNCbelow
p(not conform) above	XLSTAT_SimSPCpNCabove
p(not conform) total	XLSTAT_SimSPCpNCtotal
PPM below	XLSTAT_SimSPCPPMbelow
PPM above	XLSTAT_SimSPCPPMabove
PPM total	XLSTAT_SimSPCPPMtotal

Boîte de dialogue

OK : cliquez sur ce bouton pour créer la variable.

Annuler : cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer aucune modification.

Aide : cliquez sur ce bouton pour afficher l'aide.

 : cliquez sur ce bouton pour rétablir les options par défaut.

 : cliquez sur ce bouton pour effacer les sélections de données.

Onglet **Général**:

Nom de la variable : entrez le nom de la statistique ou sélectionnez une cellule où se trouve le nom de la variable. Si vous sélectionnez une cellule, vous pouvez spécifier si vous voulez que XLSTAT crée une référence absolue (par exemple, $A4$) ou une référence relative (par exemple, A4) à la cellule. La référence absolue ne sera pas changée si vous copiez et collez la formule XLSTAT_SimStat, contrairement ce qui serait le cas avec la référence relative.

Référence : Sélectionnez une cellule qui correspond à la variable du modèle pour laquelle vous voulez calculer la statistique.

Statistique : activez l'une des options ci-dessous puis choisissez la statistique à calculer.

- **Descriptive** : choisissez l'une des statistiques disponibles (voir la section description pour plus de détails).
- **Théorique** : choisissez l'une des statistiques disponibles (voir la section description pour plus de détails).
- **SPC** : choisissez l'une des statistiques disponibles (voir la section description pour plus de détails).

Afficher les résultats : activez cette option pour afficher les résultats détaillés concernant la statistique dans le rapport de simulation. Cette option n'est active que si vous avez sélectionné le niveau limité de filtre d'affichage « Activés » dans la boîte d'Options de XLSTAT-Sim (voir la section Options pour plus de détails).

Résultats

Un appel à la fonction XLSTAT_SimStat est généré dans la cellule active, avec les paramètres entrés. La formule suivante remplace alors le contenu de la cellule :

XLSTAT_SimStat (DistName, Reference, Statistic, Visible)

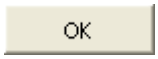
Remarque : le paramètre **Statistic** est entré sous forme d'un appel à la fonction XLSTAT correspondante. La valeur entière correspondant à l'identifiant de la statistique, difficile à interpréter, est remplacée par un l'appel à une fonction dans lequel le nom de la distribution est visible.

La couleur du fond et la couleur de la police de la cellule sont celles définies dans les options de XLSTAT-Sim.

Lancer les simulations

Une fois que le modèle de simulation a été construit en utilisant les quatre objets (distribution, variable scénario, variable résultat et statistique), vous pouvez cliquer sur le bouton  de la barre d'outils "XLSTAT-SIM". La boîte de dialogue "Lancer les simulations" apparaît alors, donnant accès à différentes options. Une fois les calculs terminés, les résultats sont affichés dans un rapport XLSTAT.

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.

 : cliquez sur ce bouton pour rétablir les options par défaut.

 : cliquez sur ce bouton pour effacer les sélections de données.

 : cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.

Onglet **Général**:

Nombre de simulations : Entrez le nombre de simulations à réaliser (valeur par défaut: 300).

Matrice de corrélation/covariance : activez cette option pour inclure dans le modèle une matrice de corrélation entre les variables aléatoires d'entrée. Les libellés des lignes et des colonnes de la matrice doivent être sélectionnés car ils sont nécessaires que pour XLSTAT sache quelles variables sont corrélées entre elles. Par conséquent, les libellés des lignes et des colonnes doivent correspondre aux noms des variables aléatoires d'entrée.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés inclus : cette option est toujours activée.

Onglet **Sorties**:

Cet onglet est subdivisé en trois sous-onglets.

Sous-Onglet Statistiques descriptives :

Statistiques descriptives : activez cette option pour calculer et afficher des statistiques descriptives pour les différentes variables. Choisissez alors les statistiques à afficher.

- **Toutes** : cliquez sur ce bouton pour tout sélectionner.
- **Aucune** : cliquez sur ce bouton pour tout désélectionner.
- **Affichage vertical** : activez cette option pour que le tableau des statistiques descriptives soit affiché verticalement (une ligne par statistique descriptive).

Quantiles : activez cette option pour afficher les quantiles des différentes variables.

Sensibilité : activez cette option pour afficher les résultats de l'analyse de sensibilité..

Détails des simulations : activez cette option pour afficher les détails des itérations de la simulation.

Sous-Onglet **Corrélations**:

Corrélations : activez cette option pour afficher les corrélations entre les variables. Si l'option "**corrélations significatives en gras** " est activée, les corrélations significatives seront affichées en gras.

- **Type de corrélation** : choisissez le type of corrélation à utiliser pour les calculs (voir la description pour plus de détails).
- **Niveau de signification (%)** : entrez le niveau de signification qui permet de déterminer si les corrélations sont significatives ou non (valeur par défaut : 5%).
- **p-values** : activez cette option pour calculer et afficher les p-values correspondant à chacune des corrélations.

Tornado : activez cette option pour afficher les résultats de l'analyse tornado.

Araignée : activez cette option pour afficher le diagramme araignée.

Onglet **SPC**:

Calculer les capacités des procédés : activez cette option pour calculer les capacités des procédés pour les variables aléatoires d'entrée, pour les variables résultat et les statistiques.

- **Nom des variables** : sélectionnez ici les données correspondant aux noms des variables pour lesquelles vous voulez calculer les capacités des procédés.
- **LSL** : veuillez entrer ici la valeur de la limite inférieure de spécification (LSL) du procédé pour chacune des variables dont le nom a été déclaré dans le champ « Nom des variables ».
- **USL** : veuillez entrer ici la valeur de la limite supérieure de spécification (USL) du procédé pour chacune des variables dont le nom a été déclaré dans le champ « Nom des variables ».
- **Cible** : activez cette option si vous voulez spécifier quelle est la valeur cible du procédé pour chacune des variables dont le nom a été déclaré dans le champ « Nom des variables ».
- **Intervalle de confiance (%)** : si le calcul des capacités du procédé sont demandés, entrez la taille en % des intervalles de confiance à calculer autour des paramètres (valeur par défaut : 95).

Onglet **Graphiques**:

Cet onglet est subdivisé en trois sous-onglets.

Sous-onglet **Histogrammes** :

Histogrammes : activez cette option pour afficher les histogrammes des échantillons. Pour la distribution théorique, la fonction de densité est affichée.

- **Barres** : choisissez cette option pour afficher des histogrammes avec une barre pour chaque intervalle.
- **Lignes continues** : choisissez cette option pour afficher des histogrammes avec une ligne continue.

Histogrammes cumulés : activez cette option pour afficher les histogrammes cumulés des échantillons.

Intervalles : choisissez l'une des options suivantes, nécessaires pour la création des histogrammes du rapport.

- **Nombre** : choisissez cette option pour entrer le nombre d'intervalles à créer.
- **Amplitude** : choisissez cette option pour définir une amplitude fixe pour les intervalles.
- **Définis par l'utilisateur** : sélectionnez une colonne contenant en ordre croissant la borne inférieure du premier intervalle, et la borne supérieure de tous les intervalles.
- **Minimum** : activez cette option pour entrer la valeur de la borne inférieure du premier intervalle. Cette valeur doit être inférieure ou égale au minimum de la série.

Sous-onglet **Box plots** :

Box plots : activez cette option pour afficher les box plots (ou graphiques boîtes et moustaches). Voir la section description pour plus de détails.

- **Horizontaux** : activez cette option pour afficher des box plots et scattergrams horizontaux.
- **Verticaux** : activez cette option pour afficher des box plots et scattergrams verticaux.
- **Grouper les graphiques** : activez cette option pour regrouper sur un même graphique les différents box plots et scattergrams de manière à pouvoir les comparer.
- **Minimum/Maximum** : activez cette option pour systématiquement afficher les points correspondant au minimum et au maximum (box plots).
- **Valeurs extrêmes** : activez cette option pour afficher les points correspondant aux valeurs extrêmes (box plots) avec un cercle évidé.
- **Position des étiquettes** : choisissez la position des étiquettes sur les graphiques verticaux. Elles peuvent être soit en bas, soit en haut, soit alternativement en bas et en haut.

Scattergrams : activez cette option pour afficher les scattergrams. La moyenne (+ rouge) et la médiane (trait rouge) sont systématiquement affichées.

Graphiques P-P (loi-normale) : activez cette option pour afficher les graphiques P-P.

Graphiques Q-Q (loi-normale) : activez cette option pour afficher les graphiques Q-Q.

Sous-onglet **Corrélations** :

Cartes des corrélations : plusieurs représentations d'une matrice des corrélations vous sont proposées.

- L'option « Echelle bleu-rouge » vous permet de représenter les corrélations faibles par des couleurs froides (bleu pour les corrélations proche de -1) et les corrélations élevées

par des couleurs chaudes (rouge pour les corrélations proches de 1).

- L'option « Noir et blanc » vous permet soit de représenter en noir les corrélations positives et en blanc les corrélations négatives (la diagonale de 1 est représentée en gris), soit de représenter en noir les corrélations significativement non nulles, et en blanc les corrélations non significativement différentes de 0.

- L'option « Motifs » vous permet de représenter les corrélations positives par des traits montant de gauche à droite, et les corrélations négatives par des traits montant de droite à gauche. Plus la corrélation est élevée en valeur absolue, plus les traits sont espacés.

Nuages de points : activez cette option pour afficher les nuages de points pour toutes les combinaisons possibles de variables deux à deux.

- **Matrice de graphiques** : activez cette option pour afficher l'ensemble des combinaisons possibles de variables deux à deux sous la forme d'un tableau à deux entrées, avec en ligne et en colonne les différentes variables.

- **Histogrammes** : activez cette option pour que XLSTAT affiche les histogrammes des variables sur la diagonale de la matrice de graphiques.

- **Q-Q plots** : activez cette option pour que XLSTAT affiche les Q-Q plots des variables sur la diagonale de la matrice de graphiques.

- **Ellipses de confiance** : activez cette option pour afficher des ellipses de confiance. Les ellipses de confiance correspondent à un intervalle de confiance à x% (x est déterminé à partir du niveau de signification spécifié dans l'onglet général) pour une loi normale bivariée de mêmes moyennes et de même matrice de covariance que les variables représentées en abscisse et en ordonnée. Cartes des corrélations : plusieurs représentations d'une matrice des corrélations vous sont proposées.

Sous-onglet **Sensibilité** :

Tornado : activez cette option pour afficher les graphiques Tornado.

Araignée : activez cette option pour afficher les graphiques araignée.

Tornado/Araignée : modifiez ici les options pour le calcul des analyses tornado et araignée.

- **Nombre de points** : choisissez le nombre de points entre les deux bornes de l'intervalle utilisé pour l'analyse tornado.

- **Valeur centrale** : choisissez l'option utilisée pour déterminer la valeur centrale des intervalles.

- **Médiane** : la valeur centrale est la médiane de la variable calculée sur l'ensemble des valeurs prises au cours des itérations.

- **Valeur par défaut de la cellule** : la valeur centrale est la valeur par défaut de la variable.

- **Définition de l'intervalle** : choisissez comment est calculé l'intervalle autour de la valeur centrale :
- **Percentile de la variable** : dans le cas où la médiane est choisie comme valeur centrale, choisissez quels deux percentiles sont à utiliser pour déterminer les bornes de l'intervalle. Vous avez le choix entre [25%, 75%], [10%, 90%] et [5%, 95%].
- **% de déviation de la valeur** : choisissez quels deux percentiles sont à utiliser pour déterminer les bornes de l'intervalle. Vous avez le choix entre [-25%, 25%], [-10%, 10%], and [-5%, 5%].

Résultats

Les premiers résultats correspondent à une synthèse sur les objets composant le modèle :

Distributions : ce tableau présente les différentes variables aléatoires d'entrée identifiées dans le projet.

Variables scénario : ce tableau présente les différentes variables scénario identifiées dans le projet.

Variable résultats : ce tableau présente les différentes variables résultat identifiées dans le projet.

Statistiques : ce tableau présente les différentes statistiques identifiées dans le projet.

Matrice de corrélation/covariance : si l'option correspondante de la boîte de dialogue est activée la matrice de corrélation ou de covariance entre les variables aléatoires d'entrée est affichée.

La section suivante du rapport présente pour chaque variable un certain nombre de résultats. Les résultats sont regroupés par type de variable (aléatoire d'entrée, résultat, statistique).

Statistiques descriptives : dans ce tableau sont affichées les statistiques choisies par l'utilisateur pour chaque variable.

Histogrammes : les histogrammes sont affichés avec pour les variables d'entrée la distribution théorique. Les détails sur les intervalles de l'histogramme sont affichés.

Sensibilité : si elle a été demandée dans la boîte de dialogue, une analyse de sensibilité est fournie pour chaque variable résultat avec notamment les corrélations, les contributions et les valeurs absolues des contributions avec les variables aléatoires d'entrée. Les contributions sont affichées sur un graphique.

Tornado : si elle a été demandée dans la boîte de dialogue, une analyse tornado est réalisée en faisant varier les variables aléatoires d'entrée et les variables scénario dans les intervalles définis. Un graphique tornado est ensuite affiché pour chaque variable résultat avec les

minimum et les maximum de la variable résultat pour chacune des variables aléatoires et scénario prises l'une après l'autre et classées par ordre décroissant di'mpact.

Graphique araignée : un tableau donnant les résultats pour chacun des points des intervalles explorés pendant l'analyse tornado est affiché. Un graphique est alors affiché sur la base de ce tableau.

Matrice de corrélation : la matrice de corrélation, le tableau des **p-values** et les graphiques associés sont affichés pour permettre de visualiser les relations entre les variables d'entrée et les variables résultat. La corrélation des cartes permettre d'identifier le potentiel des structures dans la matrice, de d'identifier rapidement les corrélations intéressantes.

Détails des simulations : un tableau avec toutes les valeurs à chaque itération et pour chaque variable du projet est affiché

Analyse de puissance

Comparer des moyennes (Puissance et taille d'échantillon)

Utilisez cet outil pour calculer la puissance ou le nombre d'observations nécessaires lors d'un test de comparaison de moyennes. XLSTAT propose aussi bien les tests t, z que des tests non paramétriques.

Dans cette section :

Description

XLSTAT propose plusieurs tests afin de comparer des moyennes. Ainsi, on peut utiliser le test t, le test z ou des tests non paramétriques tels que le test de Mann-Whitney. XLSTAT permet également d'estimer la puissance de ces tests ou de calculer le nombre d'observations nécessaires afin d'obtenir une puissance suffisante.

Lorsqu'on teste une hypothèse à l'aide d'un test statistique, on a plusieurs éléments à choisir :

- L'hypothèse nulle H0 et l'hypothèse alternative Ha
- Le test statistique à utiliser
- L'erreur de première espèce (erreur de type I) que l'on appelle aussi $alpha$. Elle se produit lorsqu'on rejette l'hypothèse nulle alors qu'elle est vraie. Elle est fixée a priori pour chaque test et vaut 5%.

L'erreur de seconde espèce ou beta est moins étudiée mais elle revêt une grande importance. En effet, elle représente la probabilité que l'on ne rejette pas l'hypothèse nulle alors qu'elle est fausse. On ne peut pas la fixer a priori mais, on peut essayer de la minimiser, en jouant sur les autres paramètres du modèle. La puissance d'un test est calculée comme 1-beta et représente la probabilité que l'on rejette l'hypothèse nulle alors qu'elle est bien fausse.

On voudra donc maximiser la puissance du test. XLSTAT permet de calculer cette puissance (ainsi que beta) lorsque les autres paramètres du test sont connus. D'autre part, il permet pour

une puissance donnée d'évaluer la taille de l'échantillon nécessaire à l'obtention de cette puissance.

Les calculs de puissance en statistique se font généralement avant que l'expérience ne soit menée. On s'en sert principalement pour estimer le nombre d'observations nécessaire pour que l'expérience ait la qualité statistique requise.

XLSTAT permet donc de comparer :

- Une moyenne à une constante (avec les tests z, t et de Wilcoxon signé)
- Deux moyennes associées à des échantillons appariés (avec les tests z, t et de Wilcoxon signé)
- Deux moyennes associées à des échantillons indépendants (avec les tests z, t et de Mann-Whitney)

On utilisera le test t lorsque la variance sur la population est estimée et le test z lorsque celle-ci est connue. Dans chaque cas, les paramètres seront différents et devront être renseignés dans la boîte de dialogue. On utilisera les tests non paramétriques lorsque les hypothèses de distribution ne pourront pas être supposées.

Méthodes

Les aides dédiées aux tests t, z et aux tests non paramétriques détaillent les méthodes en elles-mêmes.

La puissance d'un test est généralement obtenue à l'aide de la distribution non centrale associée. Ainsi, pour le test t, la distribution non centrale de Student est utilisée.

Test t pour un échantillon

La puissance de ce test est obtenue en utilisant la distribution non centrale de Student avec comme paramètre de non centralité :

$$NCP = \left| \frac{\overline{X} - X_0}{SD} \cdot \sqrt{N} \right|$$

Avec X_0 la moyenne théorique et SD l'écart-type.

La partie $\frac{\overline{X} - X_0}{SD}$ est appelé taille de l'effet (effect size), il arrive que l'on fasse varier celle-ci.

Test t pour deux échantillons appariés

La même formule que dans le cas à un seul échantillon s'applique ici aussi, mais l'écart- type se calcule de manière différente, on a donc :

$$NCP = \left| \frac{\bar{X}_1 - \bar{X}_2}{SD_{Diff}} \cdot \sqrt{N} \right|$$

avec

$$SD_{Diff} = \sqrt{(SD_1^2 + SD_2^2) - 2 \cdot Corr \cdot SD_1 \cdot SD_2}$$

et $Corr$ la corrélation entre les deux échantillons.

La partie $\frac{\bar{X}_1 - \bar{X}_2}{SD_{Diff}}$ est appelée taille de l'effet (effect size), il arrive que l'on fasse varier celle-ci.

Test t pour deux échantillons indépendants

Dans le cas de deux échantillons indépendants, l'écart-type sera calculé différemment et on utilisera la moyenne harmonique du nombre d'observations.

On a :

$$NCP = \left| \frac{\bar{X}_1 - \bar{X}_2}{SD_{Pooled}} \cdot \sqrt{\frac{N_{harmo}}{2}} \right|$$

avec

$$SD_{Pooled} = \sqrt{\frac{(N_1 - 1) \cdot SD_1^2 - (N_2 - 1) \cdot SD_2^2}{N_1 + N_2 - 2}}$$

La partie $\frac{\overline{X}_1 - \overline{X}_2}{SD_{pooled}}$ est appelée taille de l'effet (effect size), il arrive que l'on fasse varier celle-ci.

Test z pour un échantillon

Dans le cas du test z, on utilise la distribution normale classique avec un paramètre ajouté afin de décaler cette distribution.

$$NCP = \left| \frac{\overline{X} - X_0}{SD} \cdot \sqrt{N} \right|$$

Avec X_0 moyenne théorique et SD écart-type.

La partie $\frac{\overline{X} - X_0}{SD}$ est appelé taille de l'effet (effect size), il arrive que l'on fasse varier celle-ci.

Test z pour deux échantillons appariés

La même formule que dans le cas à un seul échantillon s'applique ici aussi mais l'écart- type se calcule de manière différente, on a donc :

$$NCP = \left| \frac{\bar{X}_1 - \bar{X}_2}{SD_{Diff}} \cdot \sqrt{N} \right|$$

avec

$$SD_{Diff} = \sqrt{(SD_1^2 + SD_2^2) - 2 \cdot Corr \cdot SD_1 \cdot SD_2}$$

et $Corr$ la corrélation entre les deux échantillons.

La partie $\frac{\bar{X}_1 - \bar{X}_2}{SD_{Diff}}$ est appelée taille de l'effet (effect size), il arrive que l'on fasse varier celle-ci.

Test z pour deux échantillons indépendants

Dans le cas de deux échantillons indépendants, l'écart-type sera calculé différemment et on utilisera la moyenne harmonique du nombre d'observations.

On a :

$$NCP = \left| \frac{\bar{X}_1 - \bar{X}_2}{SD_{Pooled}} \cdot \sqrt{\frac{N_{harmo}}{2}} \right|$$

avec

$$SD_{Pooled} = \sqrt{\frac{(N_1 - 1) \cdot SD_1^2 - (N_2 - 1) \cdot SD_2^2}{N_1 + N_2 - 2}}$$

La partie $\frac{\overline{X}_1 - \overline{X}_2}{SD_{pooled}}$ est appelée taille de l'effet (effect size), il arrive que l'on fasse varier celle-ci.

Tests non paramétriques

Pour les tests non paramétriques, la méthode dite ARE (*Asymptotic Relative Efficiency*) est utilisée. Cette méthode permet de relier les formules utilisées dans le cadre du test t à celle des tests non paramétriques. Elle a été introduite dans Lehmann (1975) et permet de trouver la puissance du test en utilisant un facteur ARE. Il a été montré que l'ARE minimal est de 0,864. C'est cette valeur qui est utilisée par XLSTAT. Si la distribution des données est normale, la valeur de l'ARE est de 0,955.

Pour calculer la puissance, on utilise comme distribution associée à l'hypothèse nulle, la distribution de Student t($Nk - 2$). La distribution associée à l'hypothèse alternative est la distribution de Student non centrale $t(Nk - 2, \delta)$. Le paramètre de non centralité est donné par :

$$\delta = d * \sqrt{(N_1 N_2 k)/(N_1 + N_2)}$$

où k représente l'ARE et d est la taille de l'effet tel qu'elle est définie pour le test t en fonction du type d'échantillon étudié.

L'ARE pourra être plus élevé lorsque des distributions sous-jacentes existent.

Calcul de la taille de l'échantillon

Afin de calculer le nombre d'observations nécessaires, XLSTAT utilise un algorithme de recherche de racine d'une fonction appelé algorithme Van Wijngaarden-Dekker-Brent (Brent, 1973). Cet algorithme est adapté au cas où les dérivées de la fonction ne sont pas connues. On cherche ainsi N tel que

puissance_test(N)-puissance_recherchée=0

On obtient donc la taille N telle que la puissance soit la plus proche possible de la puissance recherchée.

Taille de l'effet (effect size)

Ce concept est très important dans les calculs de puissance. En effet, Cohen (1988) a développé ce concept qui va permettre de s'affranchir d'entrer tous les paramètres du modèle (qui sont d'ailleurs souvent inconnus).

La taille de l'effet est une grandeur qui va permettre de calculer la puissance d'un test sans entrer tous les paramètres mais qui permettra de dire si l'effet des paramètres à tester est faible ou fort.

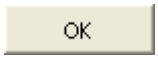
Dans le cadre des comparaisons de moyennes, les conventions de grandeurs de la taille de l'effet d sont :

- d=0,2, l'effet est faible.
- d=0,5, l'effet est modéré.
- d=0,8, l'effet est fort.

XLSTAT permet d'entrer directement la taille de l'effet.

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.

Onglet **Général**:

Objectif : choisissez entre le calcul de la puissance et la recherche du nombre d'observations (en fonction de ce choix les champs suivants seront différents).

Test : sélectionnez le test à utiliser.

Hypothèse alternative : choisissez l'hypothèse alternative que vous désirez tester.

Moyenne théorique (dans le cas d'un seul échantillon) **:** entrez la moyenne théorique à tester.

Alpha : entrez l'erreur de première espèce (elle doit être comprise entre 0,001 et 0,999).

Puissance (dans le cas où l'on recherche la taille de l'échantillon) : entrez la puissance requise (elle doit être comprise entre 0,001 et 0,999).

Taille d'échantillon (groupe 1) (dans le cas où l'on cherche à calculer la puissance) : entrez la taille de l'échantillon.

Taille d'échantillon (Groupe 2) (dans le cas où l'on cherche à calculer la puissance et qu'on a deux échantillons indépendants) : entrez la taille du second échantillon.

Rapport N1/N2 (dans le cas où l'on cherche la taille des échantillons et que l'on a deux échantillons indépendants) : entrez le rapport de la taille du premier sur celle du second échantillon.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Onglet **Taille de l'effet**:

Taille de l'effet : activez cette option afin d'entrer directement la taille de l'effet D (voir la partie description de cette aide).

Paramètres : activez cette option afin d'entrer les paramètres du test directement.

Moyenne (groupe 1) : entrez la moyenne au niveau de l'échantillon.

Moyenne (groupe 2) (dans le cas où l'on a deux échantillons) : entrez la moyenne au niveau du second échantillon.

Ecart-type (groupe 1) : Entrez l'écart type de l'échantillon.

Ecart-type (groupe 2) (dans le cas où l'on a deux échantillons) : entrez l'écart type du second échantillon.

Corrélation (dans le cas de deux échantillons appariés) : entrez la corrélation entre les deux échantillons.

Onglet **Graphiques** :

Graphiques de simulation : activez cette option si vous voulez obtenir un graphique en faisant varier différents paramètres du test. On fait varier 2 paramètres, tous les autres paramètres sont utilisés tels qu'ils ont été définis dans l'onglet général.

Axe des X : sélectionnez le paramètre à afficher sur l'axe des abscisses (X). On peut choisir entre la puissance, la taille de l'échantillon, l'erreur de première espèce (alpha) et la taille de l'effet. En fonction de ce que l'on cherche, on aura sur l'axe des Y soit la puissance, soit la taille de l'échantillon.

Taille des intervalles : sélectionnez les bornes inférieures et supérieures de l'axe des X et la taille de l'intervalle entre chaque calcul des paramètres.

Résultats

Résultats : dans ce tableau sont affichés les paramètres du test ainsi que la puissance ou le nombre d'observations nécessaires. Les paramètres obtenus par le calcul apparaissent en gras. Une phrase explicative est affichée en dessous de ce tableau.

Intervalles de simulation : ce tableau est composé de 2 colonnes, la puissance, la taille de l'échantillon ou l'alpha en fonction des paramètres sélectionnés dans la boîte de dialogue. Il permet de construire le graphique de simulation.

Graphique de simulation : Ce graphique permet de visualiser l'évolution des paramètres tel que définis dans l'onglet graphiques de la boîte de dialogue.

Exemple

Un exemple de calcul de puissance basé sur un test est disponible sur le Centre d'aide XLSTAT à l'adresse

http://www.xlstat.com/demo-pwrf.htm

Un exemple de calcul de la taille d'échantillon nécessaire est disponible sur le Centre d'aide XLSTAT à l'adresse

http://www.xlstat.com/demo-splf.htm

Bibliographie

Brent, R. P (1973) Algorithms for Minimization Without Derivatives. Englewood Cliffs, NJ: Prentice-Hall.

Cohen, J. (1988). Statistical Power Analysis for the Behavioral Sciences, Psychology Press, 2-nd Edition.

Lehmann, E. L. (1975). Nonparametrics. Statistical methods based on ranks. San Francisco, CA: Holden-Day.

Comparer des variances (Puissance et taille d'échantillon)

Utilisez cet outil pour calculer la puissance ou le nombre d'observations nécessaires lors d'un test de comparaison de variances.

Dans cette section :

Description

XLSTAT propose plusieurs tests afin de comparer des variances. XLSTAT permet également de calculer la puissance ou le nombre d'observations nécessaires pour un test basé sur la distribution F de Fisher afin de comparer des variances.

Lorsqu'on teste une hypothèse à l'aide d'un test statistique, on a plusieurs éléments à choisir :

- L'hypothèse nulle H0 et l'hypothèse alternative Ha
- Le test statistique à utiliser
- L'erreur de première espèce (erreur de type I) que l'on appelle aussi $alpha$. Elle se produit lorsqu'on rejette l'hypothèse nulle alors qu'elle est vraie. Elle est fixée a priori pour chaque test et vaut 5%.

L'erreur de seconde espèce ou beta est moins étudiée mais elle revêt une grande importance. En effet, elle représente la probabilité que l'on ne rejette pas l'hypothèse nulle alors qu'elle est fausse. On ne peut pas la fixer a priori mais, on peut essayer de la minimiser, en jouant sur les autres paramètres du modèle. La puissance d'un test est calculée comme 1-beta et représente la probabilité que l'on rejette l'hypothèse nulle alors qu'elle est bien fausse.

On voudra donc maximiser la puissance du test. XLSTAT permet de calculer cette puissance (ainsi que beta) lorsque les autres paramètres du test sont connus. D'autre part, il permet pour une puissance donnée d'évaluer la taille de l'échantillon nécessaire à l'obtention de cette puissance.

Les calculs de puissance en statistique se font généralement avant que l'expérience ne soit menée. On s'en sert principalement pour estimer le nombre d'observations nécessaire pour que l'expérience ait la qualité statistique requise.

XLSTAT permet de comparer deux variances. Les paramètres devront être renseignés dans la boîte de dialogue.

Méthodes

Les aides dédiées aux comparaisons de variances peuvent être consultées.

La puissance d'un test est généralement obtenue à l'aide de la distribution non centrale associée. Dans notre cas, nous utiliserons la distribution de Fisher.

Plusieurs hypothèses peuvent être testées, mais la plus commune est la suivante (cas bilatéral) :

- H0 : La différence entre les variances est égale à 0
- Ha : La différence entre les variances est différente de 0

Le résultat du calcul de la puissance nous permettra de donner le pourcentage d'expériences qui rejetteront l'hypothèse nulle.

Le calcul se fait en utilisant la distribution de Fisher avec comme paramètre le rapport des variances et comme degrés de libertés, les tailles des échantillons – 1.

Calcul de la taille de l'échantillon :

Afin de calculer le nombre d'observations nécessaires, XLSTAT utilise un algorithme de recherche de racine d'une fonction appelé algorithme Van Wijngaarden-Dekker-Brent (Brent, 1973). Cet algorithme est adapté au cas où les dérivées de la fonction ne sont pas connues. On cherche ainsi N tel que

puissance_test(N)-puissance_recherchée=0

On obtient donc la taille N telle que la puissance soit la plus proche possible de la puissance recherchée.

Taille de l'effet (effect size) :

Ce concept est très important dans les calculs de puissance. En effet, Cohen (1988) a développé ce concept qui va permettre de s'affranchir d'entrer tous les paramètres du modèle (qui sont d'ailleurs souvent inconnus).

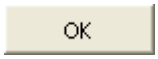
La taille de l'effet est une grandeur qui va permettre de calculer la puissance d'un test sans entrer tous les paramètres mais qui permettra de dire si l'effet des paramètres à tester est faible ou fort.

Dans le cadre des comparaisons de variances, il s'agit du rapport entre les deux variances à comparer.

XLSTAT permet d'entrer directement la taille de l'effet.

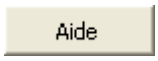
Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.



: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.

Onglet **Général**:

Objectif : choisissez entre le calcul de la puissance et la recherche du nombre d'observations (en fonction de ce choix les champs suivants seront différents).

Test : sélectionnez le test à utiliser.

Hypothèse alternative : choisissez l'hypothèse alternative que vous désirez tester.

Alpha : entrez l'erreur de première espèce (elle doit être comprise entre 0,001 et 0,999).

Puissance (dans le cas où l'on recherche la taille de l'échantillon) : entrez la puissance requise (elle doit être comprise entre 0,001 et 0,999).

Taille d'échantillon (groupe 1) (dans le cas où l'on cherche à calculer la puissance) : entrez la taille du premier échantillon.

Taille d'échantillon (groupe 2) (dans le cas où l'on cherche à calculer la puissance) : entrez la taille du second échantillon.

Rapport N1/N2 (dans le cas où l'on cherche la taille des échantillons) : entrez le rapport de la taille du premier sur celle du second échantillon.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Onglet **Taille de l'effet**:

Taille de l'effet : activez cette option afin d'entrer directement la taille de l'effet (voir la partie description de cette aide).

Paramètres : activez cette option afin d'entrer les paramètres du test directement.

Variance (groupe 1) : entrez la variance observée pour le premier échantillon.

Variance (groupe 2) : entrez la variance observée pour le second échantillon.

Onglet **Graphiques** :

Graphiques de simulation : activez cette option si vous voulez obtenir un graphique en faisant varier différents paramètres du test. On fait varier 2 paramètres, tous les autres paramètres sont utilisés tels qu'ils ont été définis dans l'onglet général.

Axe des X : sélectionnez le paramètre à afficher sur l'axe des abscisses (X). On peut choisir entre la puissance, la taille de l'échantillon, l'erreur de première espèce (alpha) et la taille de l'effet. En fonction de ce que l'on cherche, on aura sur l'axe des Y soit la puissance, soit la taille de l'échantillon.

Taille des intervalles : sélectionnez les bornes inférieures et supérieures de l'axe des X et la taille de l'intervalle entre chaque calcul des paramètres.

Résultats

Résultats : dans ce tableau sont affichés les paramètres du test ainsi que la puissance ou le nombre d'observations nécessaires. Les paramètres obtenus par le calcul apparaissent en gras. Une phrase explicative est affichée en dessous de ce tableau.

Intervalles de simulation : ce tableau est composé de 2 colonnes, la puissance, la taille de l'échantillon ou l'alpha en fonction des paramètres sélectionnés dans la boîte de dialogue. Il permet de construire le graphique de simulation.

Graphique de simulation : ce graphique permet de visualiser l'évolution des paramètres tel que définis dans l'onglet graphiques de la boîte de dialogue.

Exemple

Un exemple de calcul de puissance basé sur un test est disponible sur le Centre d'aide XLSTAT à l'adresse

http://www.xlstat.com/demo-pwrf.htm

Un exemple de calcul de la taille d'échantillon nécessaire est disponible sur le Centre d'aide XLSTAT à l'adresse

http://www.xlstat.com/demo-splf.htm

Bibliographie

Brent, R. P (1973). Algorithms for Minimization Without Derivatives. Englewood Cliffs, NJ: Prentice-Hall.

Cohen, J. (1988). Statistical Power Analysis for the Behavioral Sciences. Psychology Press, 2-nd Edition.

Comparer des proportions (Puissance et taille d'échantillon)

Utilisez cet outil pour calculer la puissance ou le nombre d'observations nécessaires lors d'un test de comparaison de proportions.

Dans cette section :

Description

XLSTAT propose plusieurs tests afin de comparer des proportions. Aussi bien des tests paramétriques que des tests non paramétriques. Ainsi on peut utiliser le test z, le test du Khi², ou encore le test du signe ou celui de McNemar. XLSTAT permet également de calculer la puissance ou le nombre d'observations nécessaire pour ces tests en utilisant soit des méthodes exactes, soit des approximations.

Lorsqu'on teste une hypothèse à l'aide d'un test statistique, on a plusieurs éléments à choisir :

- L'hypothèse nulle H0 et l'hypothèse alternative Ha
- Le test statistique à utiliser
- L'erreur de première espèce (erreur de type I) que l'on appelle aussi $alpha$. Elle se produit lorsqu'on rejette l'hypothèse nulle alors qu'elle est vraie. Elle est fixée a priori pour chaque test et vaut 5%.

L'erreur de seconde espèce ou beta est moins étudiée mais elle revêt une grande importance. En effet, elle représente la probabilité que l'on ne rejette pas l'hypothèse nulle alors qu'elle est fausse. On ne peut pas la fixer a priori mais, on peut essayer de la minimiser, en jouant sur les autres paramètres du modèle. La puissance d'un test est calculée comme 1-beta et représente la probabilité que l'on rejette l'hypothèse nulle alors qu'elle est bien fausse.

On voudra donc maximiser la puissance du test. XLSTAT permet de calculer cette puissance (ainsi que beta) lorsque les autres paramètres du test sont connus. D'autre part, il permet pour une puissance donnée d'évaluer la taille de l'échantillon nécessaire à l'obtention de cette puissance.

Les calculs de puissance en statistique se font généralement avant que l'expérience ne soit menée. On s'en sert principalement pour estimer le nombre d'observations nécessaire pour que

l'expérience ait la qualité statistique requise.

XLSTAT permet de comparer :

- Une proportion à une constante (avec le test z et différentes approximations)
- Deux proportions (avec le test z et différentes approximations)
- Des proportions dans un tableau de contingence (avec le test du Khi²)
- Des proportions dans un cadre non paramétrique (avec les tests du signe et de McNemar)

Pour chacun de ces cas, les paramètres seront différents et devront être renseignés dans la boîte de dialogue.

Méthodes

Les aides dédiées aux différents tests mentionnés plus haut détaillent les méthodes en elles-mêmes.

La puissance d'un test est généralement obtenue à l'aide de la distribution non centrale associée. Dans notre cas, nous utiliserons des approximations qui utilisent des transformations.

Comparer une proportion à une constante

L'hypothèse alternative est dans ce cas : Ha : $p_1 - p_0 \neq 0$

Différentes approximations sont alors possibles :

- Approximation en utilisant la distribution normale :

Dans ce cas, on va utiliser la distribution normale avec comme moyennes les proportions p_0 et p_1 et comme écarts-types :

$$\sqrt{\frac{p_0(1-p_0)}{N}} \text{ et } \sqrt{\frac{p_1(1-p_1)}{N}}$$

- Calcul exact en utilisant la loi binomiale de paramètres

$$\sqrt{\frac{p_0(1-p_0)}{N}} \text{ et } \sqrt{\frac{p_1(1-p_1)}{N}}$$

- Approximation en utilisant la loi béta de paramètres

$$((N-1)p_0;(N-1)(1-p_0)) \text{ et } ((N-1)p_1;(N-1)(1-p_1))$$

- Approximation en utilisant la méthode de l'arc sinus :

Cette approximation est basée sur la transformation de l'arcsin des proportions : $H(p_0)$ et $H(p_1)$. La puissance est obtenue en utilisant la distribution normale de :

$$Z_p = \sqrt{N}(H(p_0) - H(p_1)) - Z_{req}$$

Avec Z_{req} le quantile de la distribution normale pour un alpha fixé.

Comparer deux proportions

L'hypothèse alternative est dans ce cas : Ha : $p_1 - p_2 \neq 0$

Différentes approximations sont alors possibles :

- Approximation en utilisant la méthode de l'arc sinus :

Cette approximation est basée sur la transformation de l'arcsin des proportions : $H(p_1)$ et $H(p_2)$. On a donc la puissance est obtenue en utilisant la distribution normale de :

$$Z_p = \sqrt{N}(H(p_2) - H(p_1)) - Z_{req}$$

Avec Z_{req} le quantile de la distribution normale pour un alpha fixé.

- Approximation en utilisant la distribution normale :

Dans ce cas, on va utiliser la distribution normale avec comme moyennes les proportions p_1 et p_2 et comme écarts-types :

$$\sqrt{\frac{p_1(1-p_1)}{N}} \text{ et } \sqrt{\frac{p_2(1-p_2)}{N}}$$

Test du Khi²

Afin de calculer la puissance du test du Khi² dans le cas d'un tableau de contingence 2*2 (avec des proportions), on utilise la distribution non centrale du khi² avec comme paramètre de non centralité la valeur du khi² pour le tableau en question.

On cherche donc à voir si deux groupes d'observations ont les mêmes comportements par rapport à une variable binaire.

On aura :

	Groupe 1	**Groupe 2**
Positif	p_1	p_2
Négatif	$1 - p_1$	$1 - p_2$

On renseignera donc p_1, N_1 et N_2 dans la boîte de dialogue (p_2 peut être retrouvé à partir des autres paramètres car on a un seul degré de liberté).

Test du signe

Le test du signe sert à voir si la proportion de cas dans chaque groupe est égale à 50%. Il revient dans le cas de la puissance au même qu'un test sur une proportion en comparant à la valeur 0,5. On aura donc une méthode d'approximation par la loi normale ou une méthode exacte avec la loi binomiale.

On devra donc renseigner la taille de l'échantillon et la proportion dans l'un des groupes p_1 (l'autre proportion est telle que $p_2 = 1 - p_1$).

Test de McNemar

Le test de McNemar sur des proportions appariées est un cas spécifique du test sur une proportion. En effet, on peut représenter le problème avec le tableau suivant :

	Positif	**Négatif**
Positif	PP	PN
Négatif	NP	NN

On a que PP+NN+PN+NP=1. On veut essayer de voir l'effet d'un traitement, on s'intéresse donc à NP et PN, les autres valeurs n'ayant pas d'importance. On utilisera donc en entrée du test

Proportion $P1$= NP et Proportion $P2$ = PN. Avec forcément $P1 + P2 < 1$.

L'effet est donc calculé uniquement sur une proportion de NP+PN de l'échantillon. La proportion d'individus passant de positif à négatif est calculée comme NP/(NP+PN). On va donc essayer de comparer cette proportion à une valeur de 50% afin de savoir si on a plus d'individus qui vont de positif vers négatif que d'individus qui vont de négatif vers positif.

Calcul de la taille de l'échantillon

Afin de calculer le nombre d'observations nécessaires, XLSTAT utilise un algorithme de recherche de racine d'une fonction appelé algorithme Van Wijngaarden-Dekker-Brent (Brent, 1973). Cet algorithme est adapté au cas où les dérivées de la fonction ne sont pas connues. On cherche ainsi N tel que

puissance_test(N)-puissance_recherchée=0

On obtient donc la taille N telle que la puissance soit la plus proche possible de la puissance recherchée.

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.

OK : cliquez sur ce bouton pour lancer les calculs.

Annuler : cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

Aide : cliquez sur ce bouton pour afficher l'aide.

 : cliquez sur ce bouton pour rétablir les options par défaut.

 : cliquez sur ce bouton pour effacer les sélections de données.

Onglet **Général**:

Objectif : choisissez entre le calcul de la puissance et la recherche du nombre d'observations (en fonction de ce choix les champs suivants seront différents).

Test : sélectionnez le test à utiliser.

Hypothèse alternative : choisissez l'hypothèse alternative que vous désirez tester.

Proportion théorique (dans le cas d'un seul échantillon) : entrez la proportion théorique à tester.

Alpha : entrez l'erreur de première espèce (elle doit être comprise entre 0,001 et 0,999).

Puissance (dans le cas où l'on recherche la taille de l'échantillon) **:** entrez la puissance requise (elle doit être comprise entre 0,001 et 0,999).

Taille d'échantillon (groupe1) (dans le cas où l'on cherche à calculer la puissance) : entrez la taille de l'échantillon.

Taille d'échantillon (groupe 2) (dans le cas où l'on cherche à calculer la puissance et qu'on a deux échantillons indépendants) **:** entrez la taille du second échantillon.

Rapport N1/N2 (dans le cas où l'on cherche la taille des échantillons et que l'on a deux échantillons indépendants) **:** entre le rapport de la taille du premier sur celle du second échantillon.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Onglet **Taille de l'effet** :

Proportion (groupe 1) : entrez la proportion observée pour l'échantillon (du premier dans le cas où l'on en a deux). Pour les tests du khi², de McNemar et du signe, se référer à la partie

description de ce chapitre.

Proportion (groupe 2) (dans le cas où l'on a deux échantillons) **:** entrez la proportion observée pour le second échantillon. Pour les tests du khi², de McNemar et du signe, se référer à la partie description de ce chapitre.

Onglet **Graphiques** :

Graphiques de simulation : activez cette option si vous voulez obtenir un graphique en faisant varier différents paramètres du test. On fait varier 2 paramètres, tous les autres paramètres sont utilisés tels qu'ils ont été définis dans l'onglet général.

Axe des X : sélectionnez le paramètre à afficher sur l'axe des abscisses (X). On peut choisir entre la puissance, la taille de l'échantillon, l'erreur de première espèce (alpha) et la taille de l'effet. En fonction de ce que l'on cherche, on aura sur l'axe des Y soit la puissance, soit la taille de l'échantillon.

Taille des intervalles : sélectionnez les bornes inférieures et supérieures de l'axe des X et la taille de l'intervalle entre chaque calcul des paramètres.

Résultats

Résultats : dans ce tableau sont affichés les paramètres du test ainsi que la puissance ou le nombre d'observations nécessaires. Les paramètres obtenus par le calcul apparaissent en gras. Une phrase explicative est affichée en dessous de ce tableau.

Intervalles de simulation : ce tableau est composé de 2 colonnes, la puissance, la taille de l'échantillon ou l'alpha en fonction des paramètres sélectionnés dans la boîte de dialogue. Il permet de construire le graphique de simulation.

Graphique de simulation : ce graphique permet de visualiser l'évolution des paramètres tel que définis dans l'onglet graphiques de la boîte de dialogue.

Exemple

Un exemple de calcul de puissance basé sur un test est disponible sur le Centre d'aide XLSTAT à l'adresse

http://www.xlstat.com/demo-pwrf.htm

Un exemple de calcul de la taille d'échantillon nécessaire est disponible sur le Centre d'aide XLSTAT à l'adresse

http://www.xlstat.com/demo-splf.htm

Bibliographie

Brent, R. P (1973). Algorithms for Minimization Without Derivatives. Englewood Cliffs, NJ: Prentice-Hall.

Cohen, J. (1988). Statistical Power Analysis for the Behavioral Sciences. Psychology Press, 2^{nd} Edition.

Comparer des corrélations (Puissance et taille d'échantillon)

Utilisez cet outil pour calculer la puissance ou le nombre d'observations nécessaires lors d'un test statistique afin de comparer des corrélations de Pearson.

Dans cette section :

Description

XLSTAT propose un test afin de comparer des corrélations. XLSTAT permet également de calculer la puissance ou le nombre d'observations nécessaire pour ce test.

Lorsqu'on teste une hypothèse à l'aide d'un test statistique, on a plusieurs éléments à choisir :

- L'hypothèse nulle H0 et l'hypothèse alternative Ha
- Le test statistique à utiliser
- L'erreur de première espèce (erreur de type I) que l'on appelle aussi $alpha$. Elle se produit lorsqu'on rejette l'hypothèse nulle alors qu'elle est vraie. Elle est fixée a priori pour chaque test et vaut 5%.

L'erreur de seconde espèce ou beta est moins étudiée mais elle revêt une grande importance. En effet, elle représente la probabilité que l'on ne rejette pas l'hypothèse nulle alors qu'elle est fausse. On ne peut pas la fixer a priori mais, on peut essayer de la minimiser, en jouant sur les autres paramètres du modèle. La puissance d'un test est calculée comme 1-beta et représente la probabilité que l'on rejette l'hypothèse nulle alors qu'elle est bien fausse.

On voudra donc maximiser la puissance du test. XLSTAT permet de calculer cette puissance (ainsi que beta) lorsque les autres paramètres du test sont connus. D'autre part, il permet pour une puissance donnée d'évaluer la taille de l'échantillon nécessaire à l'obtention de cette puissance.

Les calculs de puissance en statistique se font généralement avant que l'expérience ne soit menée. On s'en sert principalement pour estimer le nombre d'observations nécessaire pour que l'expérience ait la qualité statistique requise.

XLSTAT permet donc de comparer :

- Une corrélation à 0
- Une corrélation à une constante
- Deux corrélations

Pour chacun de ces cas, les paramètres seront différents et devront être renseignés dans la boîte de dialogue.

Méthodes

Les aides dédiées aux tests énoncés plus haut détaillent les méthodes en elles-mêmes.

La puissance d'un test est généralement obtenue à l'aide de la distribution non centrale associée. Dans notre cas, nous utiliserons la distribution non centrale de Student ou des approximations en utilisant la loi normale.

Comparaison d'une corrélation à 0

L'hypothèse alternative est dans ce cas : Ha : $r \neq 0$

La méthode utilisée est une méthode exacte basée sur la distribution non centrale de Student.

Le paramètre de non centralité utilisé est le suivant :

$$NCP = \sqrt{\frac{r^2}{1-r^2}} \cdot \sqrt{N}$$

La partie $\frac{r^2}{1-r^2}$ est appelé taille de l'effet (effect size), il arrive que l'on fasse varier celle-ci.

Comparaison d'une corrélation à une constante

L'hypothèse alternative est dans ce cas : Ha : $r \neq r_0$

Le calcul de la puissance se fait en utilisant une approximation par la loi normale. On utilise la transformation Z de Fisher :

$$Z_r = \frac{1}{2}\log(\frac{1+r}{1-r})$$

On prend comme taille de l'effet :

$$Q = |Z_r - Z_{r_0}|$$

La puissance est alors trouvée en utilisant l'aire sous la courbe de la distribution normale à gauche de Z_p:

$$Z_p = Q \cdot \sqrt{N-3} - Z_{req}$$

Avec Z_{req} le quantile de la distribution normale pour un alpha fixé.

Comparaison de deux corrélations

L'hypothèse alternative est dans ce cas : Ha : $r_1 - r_2 \neq 0$

Le calcul de la puissance se fait en utilisant une approximation par la loi normale. On utilise la transformation Z de Fisher :

$$Z_r = \frac{1}{2}\log(\frac{1+r}{1-r})$$

On prend comme effet :

$$Q = |Z_{r_1} - Z_{r_2}|$$

La puissance est alors trouvée en utilisant l'aire sous la courbe de la distribution normale à gauche de Z_p:

$$Z_p = Q \cdot \sqrt{\frac{N' - 3}{2}} - Z_{req}$$

Avec Z_{req} le quantile de la distribution normale pour un alpha fixé et

$$N' = \frac{2(N_1 - 3)(N_2 - 3)}{N_1 + N_2 - 6} + 3$$

Calcul de la taille de l'échantillon

Afin de calculer le nombre d'observations nécessaires, XLSTAT utilise un algorithme de recherche de racine d'une fonction appelé algorithme Van Wijngaarden-Dekker-Brent (Brent, 1973). Cet algorithme est adapté au cas où les dérivées de la fonction ne sont pas connues. On cherche ainsi N tel que

puissance(N)-puissance_recherchée=0

On obtient donc la taille N telle que la puissance soit la plus proche possible de la puissance recherchée.

Taille de l'effet (effect size)

Ce concept est très important dans les calculs de puissance. En effet, Cohen (1988) a développé ce concept qui va permettre de s'affranchir d'entrer tous les paramètres du modèle (qui sont d'ailleurs souvent inconnus).

La taille de l'effet est une grandeur qui va permettre de calculer la puissance d'un test sans entrer tous les paramètres mais qui permettra de dire si l'effet des paramètres à tester est faible ou fort.

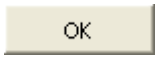
Dans le cadre des comparaisons de corrélations, les conventions de grandeurs de la taille de l'effet q sont :

- Q=0,1, l'effet est faible.
- Q=0,3, l'effet est modéré.
- Q=0,5, l'effet est fort.

XLSTAT permet d'entrer directement la taille de l'effet.

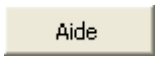
Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.



: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.

Onglet **Général**:

Objectif : choisissez entre le calcul de la puissance et la recherche du nombre d'observations (en fonction de ce choix les champs suivants seront différents).

Test : sélectionnez le test à utiliser.

Hypothèse alternative : choisissez l'hypothèse alternative que vous désirez tester.

Corrélation théorique (dans le cas d'un seul échantillon) : entrez la corrélation théorique à tester.

Alpha : entrez l'erreur de première espèce (elle doit être comprise entre 0,001 et 0,999).

Puissance (dans le cas où l'on recherche la taille de l'échantillon) : entrez la puissance requise (elle doit être comprise entre 0,001 et 0,999).

Taille d'échantillon (groupe 1) (dans le cas où l'on cherche à calculer la puissance) : entrez la taille de l'échantillon.

Taille d'échantillon (groupe 2) (dans le cas où l'on cherche à calculer la puissance et qu'on a deux échantillons indépendants) : entrez la taille du second échantillon.

Rapport N1/N2 (dans le cas où l'on cherche la taille des échantillons et que l'on a deux échantillons indépendants) : entrez le rapport de la taille du premier sur celle du second échantillon.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Onglet **Taille de l'effet**:

Taille de l'effet : activez cette option afin d'entrer directement la taille de l'effet D (voir la partie description de cette aide).

Paramètres : activez cette option afin d'entrer les paramètres du test directement.

Corrélation (groupe 1) : entrez la corrélation observée pour l'échantillon (du premier dans le cas où l'on en a deux).

Corrélation (groupe 2) (dans le cas où l'on a deux échantillons) : entrez la corrélation observée pour le second échantillon.

Onglet **Graphiques** :

Graphiques de simulation : activez cette option si vous voulez obtenir un graphique en faisant varier différents paramètres du test. On fait varier 2 paramètres, tous les autres paramètres sont utilisés tels qu'ils ont été définis dans l'onglet général.

Axe des X : sélectionnez le paramètre à afficher sur l'axe des abscisses (X). On peut choisir entre la puissance, la taille de l'échantillon, l'erreur de première espèce (alpha) et la taille de l'effet. En fonction de ce que l'on cherche, on aura sur l'axe des Y soit la puissance, soit la taille de l'échantillon.

Taille des intervalles : sélectionnez les bornes inférieures et supérieures de l'axe des X et la taille de l'intervalle entre chaque calcul des paramètres.

Résultats

Résultats : dans ce tableau sont affichés les paramètres du test ainsi que la puissance ou le nombre d'observations nécessaires. Les paramètres obtenus par le calcul apparaissent en gras. Une phrase explicative est affichée en dessous de ce tableau.

Intervalles de simulation : ce tableau est composé de 2 colonnes, la puissance, la taille de l'échantillon ou l'alpha en fonction des paramètres sélectionnés dans la boîte de dialogue. Il permet de construire le graphique de simulation.

Graphique de simulation : ce graphique permet de visualiser l'évolution des paramètres tel que définis dans l'onglet graphiques de la boîte de dialogue.

Exemple

Un exemple de calcul de puissance basé sur un test est disponible sur le Centre d'aide XLSTAT à l'adresse

http://www.xlstat.com/demo-pwrf.htm

Un exemple de calcul de la taille d'échantillon nécessaire est disponible sur le Centre d'aide XLSTAT à l'adresse

http://www.xlstat.com/demo-splf.htm

Bibliographie

Brent R. P (1973). Algorithms for Minimization Without Derivatives. Englewood Cliffs, NJ: Prentice-Hall.

Cohen J. (1988). Statistical Power Analysis for the Behavioral Sciences. Psychology Press, 2-nd Edition.

Régression linéaire (Puissance et taille d'échantillon)

Utilisez cet outil pour calculer la puissance ou le nombre d'observations nécessaires lors d'une régression linéaire simple ou multiple.

Dans cette section :

Description

XLSTAT propose un outil permettant d'appliquer un modèle de régression linéaire. XLSTAT permet également d'estimer la puissance ou de calculer le nombre d'observations nécessaires associée aux variations du R^2 dans le cadre d'une régression linéaire.

Lorsqu'on teste une hypothèse à l'aide d'un test statistique, on a plusieurs éléments à choisir :

- L'hypothèse nulle H0 et l'hypothèse alternative Ha
- Le test statistique à utiliser
- L'erreur de première espèce (erreur de type I) que l'on appelle aussi $alpha$. Elle se produit lorsqu'on rejette l'hypothèse nulle alors qu'elle est vraie. Elle est fixée a priori pour chaque test et vaut 5%.

L'erreur de seconde espèce ou beta est moins étudiée mais elle revêt une grande importance. En effet, elle représente la probabilité que l'on ne rejette pas l'hypothèse nulle alors qu'elle est fausse. On ne peut pas la fixer a priori mais, on peut essayer de la minimiser, en jouant sur les autres paramètres du modèle. La puissance d'un test est calculée comme 1-beta et représente la probabilité que l'on rejette l'hypothèse nulle alors qu'elle est bien fausse.

On voudra donc maximiser la puissance du test. XLSTAT permet de calculer cette puissance (ainsi que beta) lorsque les autres paramètres du test sont connus. D'autre part, il permet pour une puissance donnée d'évaluer la taille de l'échantillon nécessaire à l'obtention de cette puissance.

Les calculs de puissance en statistique se font généralement avant que l'expérience ne soit menée. On s'en sert principalement pour estimer le nombre d'observations nécessaire pour que l'expérience ait la qualité statistique requise.

XLSTAT permet donc de comparer :

- La valeur du R^2 à 0
- L'augmentation du R^2 lorsqu'on ajoute de nouvelles variables explicatives au modèle.

Ceci revient à tester les hypothèses suivantes :

- H0 : $R^2 = 0$ / Ha : $R^2 \neq 0$
- H0 : L'augmentation du R^2 est égale à 0 / Ha : L'augmentation du R^2 est différente de 0.

Taille de l'effet (effect size)

Ce concept est très important dans les calculs de puissance. En effet, Cohen (1988) a développé ce concept qui va permettre de s'affranchir d'entrer tous les paramètres du modèle (qui sont d'ailleurs souvent inconnus).

La taille de l'effet est une grandeur qui permet de calculer la puissance d'un test sans entrer tous les paramètres mais qui permet de dire si l'effet des paramètres à tester est faible ou fort.

Dans le cadre de la régression linéaire, les conventions de grandeurs de la taille de l'effet f^2 sont :

- $f^2 = 0,02$, l'effet est faible.
- $f^2 = 0,1$, l'effet est modéré.
- $f^2 = 0,35$, l'effet est fort.

XLSTAT permet d'entrer directement la taille de l'effet, mais permet aussi d'entrer des paramètres du modèle qui permettront de calculer la taille de l'effet. Nous en détaillons les calculs ci-dessous :

- En utilisant les variances : On peut utiliser les variances du modèle afin de définir la taille de l'effet. En prenant Var_{exp} pour la variance expliquée par les variables explicatives que l'on désire tester et Var_{error} pour la variance de l'erreur ou variance résiduelle du modèle, on aura :

$$f^2 = \frac{var_{exp}}{var_{error}}$$

- En utilisant le R^2 (dans le cas de l'hypothèse nulle $R^2 = 0$) : On entre alors la valeur estimée du carré de la corrélation multiple théorique (ρ^2) pour le modèle analysé. On aura :

$$f^2 = \frac{\rho^2}{1 - \rho^2}$$

- En utilisant le R^2 partiel (dans le cas du test sur l'augmentation du R^2) : On entre alors la valeur du R^2 partiel qui représente l'augmentation du R^2 lorsqu'on ajoute un groupe de variables. On aura :

$$f^2 = \frac{R_{part}2}{1 - R_{part}2}$$

- En utilisant les corrélations entre les variables du modèle (dans le cas du test R^2 différent de 0) : On doit alors sélectionner un vecteur contenant les corrélations entre les variables explicatives et la variable dépendante $Corr_Y$ et une matrice carrée contenant les corrélations entre les variables explicatives du modèle $Corr_X$. On aura :

$$f^2 = \frac{Corr_Y^t(Corr_X^{-1})Corr_Y}{1 - Corr_Y^t(Corr_X^{-1})Corr_Y}$$

Une fois la taille de l'effet définie, on peut calculer la taille de l'échantillon nécessaire ou la puissance obtenue.

Méthodes

L'aide dédiée à la régression linéaire décrit en détails cette méthode.

La puissance d'un test est généralement obtenue à l'aide de la distribution non centrale associée. Ainsi, pour le cas de la régression linéaire, la distribution non centrale de Fisher est utilisée.

La puissance de ce test est obtenue en utilisant la distribution non centrale de Fisher avec comme degrés de libertés : DL1 est le nombre de variables explicatives testées, DL2 est la taille de l'échantillon à laquelle on soustrait le nombre total de variables explicatives inclus dans le modèle plus un et comme paramètre de non centralité :

$$NCP = f^2N$$

Calcul de la taille de l'échantillon

Afin de calculer le nombre d'observations nécessaires, XLSTAT utilise un algorithme de recherche de racine d'une fonction appelée algorithme Van Wijngaarden-Dekker-Brent (Brent, 1973). Cet algorithme est adapté au cas où les dérivées de la fonction ne sont pas connues. On cherche ainsi N tel que

puissance_test(N)-puissance_recherchée=0

On obtient donc la taille N telle que la puissance soit la plus proche possible de la puissance recherchée.

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.

OK : cliquez sur ce bouton pour lancer les calculs.

Annuler : cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

Aide : cliquez sur ce bouton pour afficher l'aide.

 : cliquez sur ce bouton pour rétablir les options par défaut.

 : cliquez sur ce bouton pour effacer les sélections de données.

Onglet **Général**:

Objectif : choisissez entre le calcul de la puissance et la recherche du nombre d'observations (en fonction de ce choix les champs suivants seront différents).

Test : sélectionnez le test à utiliser suivant que l'on veut vérifier que le R^2 est significativement différent de 0 ou que l'ajout de nouvelles variables a un effet significatif sur le R^2.

Alpha : entrez l'erreur de première espèce (elle doit être comprise entre 0,001 et 0,999).

Puissance (dans le cas où l'on recherche la taille de l'échantillon) : entrez la puissance requise (elle doit être comprise entre 0,001 et 0,999).

Taille d'échantillon (dans le cas où l'on cherche à calculer la puissance) : entrez la taille de l'échantillon.

Nombre de prédicteurs testés : entrez le nombre de variables explicatives que l'on souhaite analyser dans le modèle.

Nombre total de prédicteurs (dans le cas où l'on teste l'augmentation du R^2 suite à l'ajout de nouvelles variables) : entrez le nombre total de variables explicatives inclues dans le modèle.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Onglet **Taille de l'effet**:

Recherche de la taille de l'effet : sélectionnez la méthode à utiliser afin de calculer la taille de l'effet (soit en entrant directement la taille de l'effet, soit à partir des variances du modèle, soit à

partir du R^2 directement, soit en utilisant les corrélations entre les variables explicatives).

Talle de l'effet f^2 (dans le cas où l'on entre directement la taille de l'effet) : entrez la taille de l'effet désirée (voir la partie description pour connaître les ordres de grandeur).

Variance expliquée (dans le cas où l'on part des variances pour définir la taille de l'effet) : Entrez la valeur de la variance expliquée par les variables explicatives étudiées.

Variance de l'erreur (dans le cas où l'on part des variances pour définir la taille de l'effet) : entrez la valeur de la variance de l'erreur du modèle complet.

R^2 partiel (dans le cas où l'on utilise la méthode directe afin de calculer la taille de l'effet) : entrez la valeur du R^2 partiel, c'est-à-dire de la modification du R^2 lorsqu'on ajoute un groupe de variables explicatives au modèle.

rho² (dans le cas où l'on utilise la méthode à partir du R^2 afin de calculer la taille de l'effet) : entrez la valeur du R^2 théorique du modèle.

Les champs suivant apparaîssent lorsqu'on sélectionne comme test, R^2 différent de 0 et comme méthode pour calculer la taille de l'effet, les corrélations entre prédicteurs.

Corrélations avec les Y : sélectionnez une colonne de valeur correspondant aux corrélations entre les variables explicatives et la variable réponse Y du modèle. Ce vecteur doit avoir un nombre de lignes égal au nombre de variables explicatives. Il ne faut pas sélectionner le libellé de la colonne mais uniquement les valeurs.

Corrélations entre les prédicteurs : sélectionnez un tableau de valeur correspondant aux corrélations entre les variables explicatives du modèle. Ce tableau doit être symétrique, avoir des 1 sur la diagonale et avoir un nombre de lignes et de colonnes égal au nombre de variables explicatives. Il ne faut pas sélectionner le libellé des colonnes ni des lignes mais uniquement les valeurs.

Onglet **Graphiques** :

Graphiques de simulation : activez cette option si vous voulez obtenir un graphique en faisant varier différents paramètres du test. On fait varier 2 paramètres, tous les autres paramètres sont utilisés tels qu'ils ont été définis dans l'onglet général.

Axe des X : sélectionnez le paramètre à afficher sur l'axe des abscisses (X). On peut choisir entre la puissance, la taille de l'échantillon, l'erreur de première espèce (alpha) et la taille de l'effet. En fonction de ce que l'on cherche, on aura sur l'axe des Y soit la puissance, soit la taille de l'échantillon.

Taille des intervalles : sélectionnez les bornes inférieures et supérieures de l'axe des X et la taille de l'intervalle entre chaque calcul des paramètres.

Résultats

Entrées : dans ce tableau sont affichées les paramètres ayant permis de calculer la puissance ou le nombre d'observations nécessaires.

Résultats : dans ce tableau sont affichées l'alpha, la taille de l'effet ainsi que la puissance et le nombre d'observations nécessaires. Les paramètres obtenus par le calcul apparaissent en gras. Une phrase explicative est affichée en dessous de ce tableau.

Intervalles de simulation : ce tableau est composé de 2 colonnes, la puissance, la taille de l'échantillon ou l'alpha en fonction des paramètres sélectionnés dans la boîte de dialogue. Il permet de construire le graphique de simulation.

Graphique de simulation : Ce graphique permet de visualiser l'évolution des paramètres tel que définis dans l'onglet graphiques de la boîte de dialogue.

Exemple

Un exemple de calcul de puissance basé sur un test est disponible sur le Centre d'aide XLSTAT à l'adresse

http://www.xlstat.com/demo-pwrf.htm

Un exemple de calcul de la taille d'échantillon nécessaire est disponible sur le Centre d'aide XLSTAT à l'adresse

http://www.xlstat.com/demo-splf.htm

Bibliographie

Brent R. P (1973). Algorithms for Minimization Without Derivatives. Englewood Cliffs, NJ: Prentice-Hall.

Cohen J. (1988). Statistical Power Analysis for the Behavioral Sciences, Psychology Press, 2-nd Edition.

Dempster A.P. (1969). Elements of Continuous Multivariate Analysis. Addison-Wesley, Reading.

ANOVA/ANCOVA (Puissance et taille d'échantillon)

Utilisez cet outil pour calculer la puissance ou le nombre d'observations nécessaires lors d'une analyse de la variance avec ou sans mesures répétées, ou lors d'une analyse de la covariance.

Dans cette section :

Description

XLSTAT propose des outils permettant d'appliquer un modèle d'analyse de la variance, d'analyse de la variance à mesures répétées et d'analyse de la covariance. XLSTAT permet également d'estimer la puissance ou de calculer le nombre d'observations nécessaires dans le cadre de ces méthodes.

Lorsqu'on teste une hypothèse à l'aide d'un test statistique, on a plusieurs éléments à choisir :

- L'hypothèse nulle H0 et l'hypothèse alternative Ha
- Le test statistique à utiliser
- L'erreur de première espèce (erreur de type I) que l'on appelle aussi $alpha$. Elle se produit lorsqu'on rejette l'hypothèse nulle alors qu'elle est vraie. Elle est fixée a priori pour chaque test et vaut 5%.

L'erreur de seconde espèce ou beta est moins étudiée mais elle revêt une grande importance. En effet, elle représente la probabilité que l'on ne rejette pas l'hypothèse nulle alors qu'elle est fausse. On ne peut pas la fixer a priori mais, on peut essayer de la minimiser, en jouant sur les autres paramètres du modèle. La puissance d'un test est calculée comme 1-beta et représente la probabilité que l'on rejette l'hypothèse nulle alors qu'elle est bien fausse.

On voudra donc maximiser la puissance du test. XLSTAT permet de calculer cette puissance (ainsi que beta) lorsque les autres paramètres du test sont connus. D'autre part, il permet pour une puissance donnée d'évaluer la taille de l'échantillon nécessaire à l'obtention de cette puissance.

Les calculs de puissance en statistique se font généralement avant que l'expérience ne soit menée. On s'en sert principalement pour estimer le nombre d'observations nécessaire pour que l'expérience ait la qualité statistique requise.

XLSTAT permet donc de tester :

- Dans le cas de l'ANOVA à un facteur ou à plusieurs facteurs et interactions, ainsi que dans le cas de l'ANCOVA :
- H0 : les moyennes des groupes associées au facteur testé sont toutes égales.
- Ha : il existe au moins une moyenne qui est différente des autres.
- Dans le cas d'un facteur intra-sujets pour une ANOVA à mesures répétées :
- H0 : les moyennes des groupes associées au facteur intra-sujet sont toutes égales.
- Ha : il existe au moins une moyenne qui est différente des autres.
- Dans le cas d'un facteur inter-sujets pour une ANOVA à mesures répétées :
- H0 : les moyennes des groupes associées au facteur inter-sujets sont toutes égales.
- Ha : il existe au moins une moyenne qui est différente des autres.
- Dans le cas d'une interaction entre un facteur intra-sujets et un facteur inter-sujets pour une ANOVA à mesures répétées :
- H0 : les moyennes des groupes associées à l'interaction intra-inter sont toutes égales.
- Ha : il existe au moins une moyenne qui est différente des autres.

Taille de l'effet (effect size)

Ce concept est très important dans les calculs de puissance. En effet, Cohen (1988) a développé ce concept qui va permettre de s'affranchir d'entrer tous les paramètres du modèle (qui sont d'ailleurs souvent inconnus).

La taille de l'effet est une grandeur qui va permettre de calculer la puissance d'un test sans entrer tous les paramètres mais qui permettra de dire si l'effet des paramètres à tester est faible ou fort.

Dans le cadre de l'ANOVA et de l'ANCOVA, les conventions de grandeurs de la taille de l'effet f sont :

- $f = 0.1$, l'effet est faible.
- $f = 0.25$, l'effet est modéré.
- $f = 0.4$, l'effet est fort.

XLSTAT permet d'entrer directement la taille de l'effet, mais permet aussi d'entrer des paramètres du modèle qui permettront de calculer la taille de l'effet. Nous en détaillons les calculs ci-dessous :

- En utilisant les variances : On peut utiliser les variances du modèle afin de définir la taille de l'effet. En prenant Var_{exp} pour la variance expliquée par les facteurs que l'on désire tester et Var_{error} pour la variance de l'erreur ou variance résiduelle du modèle, on aura :

$$f^2 = \frac{var_{exp}}{var_{error}}$$

- En utilisant la méthode directe : On entre alors la valeur estimée de η^2, qui est le rapport de la variance expliquée par la variance totale du modèle. Pour plus de détails sur cet indice on se référera à Cohen (1988) (chap.8.2). On aura :

$$f = \sqrt{\frac{\eta^2}{1-\eta^2}}$$

- En utilisant les moyennes pour chaque groupe (uniquement dans le cas d'une ANOVA à un facteur ou d'une analyse inter-sujets d'une ANOVA à mesures répétées) : On sélectionne alors un vecteur comprenant les moyennes calculées pour chaque groupe. Il est aussi possible d'avoir des groupes de tailles différentes, dans ce cas, il faut aussi sélectionner un vecteur avec les différentes tailles (l'option classique suppose que tous les groupes ont des tailles égales). On aura :

$$f = \frac{\sqrt{\sum_i \frac{(mi-m)^2}{k}}}{SD_{intra}}$$

Avec m_i moyenne pour le groupe i, m la moyenne des moyennes, k nombre de groupes et SD_{intra} l'écart-type intra-groupe.

- Lorsqu'on effectue une analyse de la covariance (ANCOVA), il faut ajouter un terme dans le calcul de la taille de l'effet, on multiplie f par :

$$\sqrt{\frac{1}{1-\rho^2}}$$

où ρ^2 est la valeur théorique du carré de la corrélation multiple des variables explicative quantitative du modèle.

Une fois la taille de l'effet définie, on peut calculer la taille de l'échantillon nécessaire ou la puissance obtenue.

Méthodes

Les aides dédiées aux différentes méthodes donnent de plus amples détails sur chacune d'entre elles.

Nous introduisons les notations suivantes :

- NbrGroup : Nombre de groupes que l'on désire tester.
- N : taille de l'échantillon
- NumeratorDF : Degré de liberté du numérateur, voir plus bas pour une explication détaillée.
- NbrRep : Nombre de répétitions (ou de mesures) dans le cadre d'une ANOVA à mesures répétées.
- ρ : Corrélation entre les mesures dans le cadre d'une ANOVA à mesures répétées.
- ϵ: Correction pour la non-sphéricité de Geisser-Greenhouse.
- NbrPred: Nombre de variables explicatives quantitatives dans le cadre d'une ANCOVA.

La puissance d'un test est généralement obtenue à l'aide de la distribution non centrale associée. Ainsi, pour le cas de l'ANOVA, la distribution non centrale de Fisher est utilisée.

Pour chaque méthode, on donnera les premiers et seconds degrés de liberté ainsi que le paramètre de non centralité :

- ANOVA à un effet :

$$DF1 = \text{NbrGroup } - 1 \quad DF2 = N - NbrGroup \quad NCP = f^2 N$$

- ANOVA effets fixes et interactions :

$$DF1 = NumeratorDF \quad DF2 = N - NbrGroup \quad NCP = f^2 N$$

- ANOVA à mesures répétées intra facteur :

$$DF1 = NbrRep - 1 \quad DF2 = (N - NbrGroup)(NbrRep - 1) \quad NCP = f^2 \frac{N.NbrRep}{1-\rho}\epsilon$$

- ANOVA à mesures répétées inter facteurs :

$$DF1 = NbrGroup - 1 \quad DF2 = N - NbrGroup \quad NCP = f^2 \frac{N.NbrRep}{1-\rho(NbrRep-1)}\epsilon$$

- ANOVA à mesures répétées intra-inter facteurs :

$$DF1 = (NbrRep - 1)(NbrGroup - 1) \quad DF2 = (N - NbrGroup)(NbrRep - 1) \quad NCP =$$

- ANCOVA :

$$DF1 = NumeratorDF \quad DF2 = N - NbrGroup - NbrPredictors \quad NCP = f^2 N$$

Calcul de la taille de l'échantillon

Afin de calculer le nombre d'observations nécessaires, XLSTAT utilise un algorithme de recherche de racine d'une fonction appelé algorithme Van Wijngaarden-Dekker-Brent (Brent, 1973). Cet algorithme est adapté au cas où les dérivées de la fonction ne sont pas connues. On cherche ainsi N tel que

puissance_test(N)-puissance_recherchée=0

On obtient donc la taille N telle que la puissance soit la plus proche possible de la puissance recherchée.

Calcul des degrés de liberté du numérateur

Dans le cadre d'une ANOVA à facteurs fixes et interactions ou d'une ANCOVA, XLSTAT propose de renseigner le nombre de degrés de liberté du numérateur pour la loi de Fisher non centrale.

Supposons que nous avons un modèle à 3 facteurs, A (2 groupes), B (3 groupes), C (3 groupes), 3 interactions de niveau 2 AB, AC et BC et 1 interaction de niveau 3, ABC. On aura donc 18 groupes.

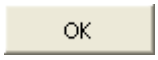
Pour tester les effets principaux, par exemple A, nous avons : $NbGroups = 18$ et $NumeratorDF = (2-1) = 1$.

Pour tester les interactions, par exemple AB, nous avons $NbGroups = 18$ et $NumeratorDF = (2-1)(3-1) = 2$. Si on désire tester l'interaction de niveau 3 (ABC), on prendra $NbGroups = 18$ et $NumeratorDF = (2-1)(3-1)(3-1) = 4$.

Dans le cadre d'une ANCOVA, les calculs seront similaires.

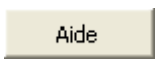
Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.



: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.

Onglet **Général**:

Objectif : choisissez entre le calcul de la puissance et la recherche du nombre d'observations (en fonction de ce choix les champs suivants seront différents).

Test : sélectionnez la méthode que vous désirez utiliser. Voir la partie description de cette aide pour plus de détails).

Alpha : entrez l'erreur de première espèce (elle doit être comprise entre 0,001 et 0,999).

Puissance (dans le cas où l'on recherche la taille de l'échantillon) : entrez la puissance requise (elle doit être comprise entre 0,001 et 0,999).

Taille d'échantillon (dans le cas où l'on cherche à calculer la puissance) : entrez la taille de l'échantillon.

Nombre de groupes : entrez le nombre total de groupes inclus dans le modèle.

NumDDL (uniquement dans le cas de l'ANOVA à plus d'un facteur et de l'ANCOVA) : entrez le degré de liberté du numérateur de la loi F. Celui-ci dépendra du type de facteur testé (pour plus de détails, voir la partie description de cette aide).

Nombre de mesures (uniquement dans le cas de l'ANOVA à mesures répétées : entrez le nombre de mesures (répétitions) de l'ANOVA à mesures répétées.

Corrélation entre les mesures (uniquement dans le cas de l'ANOVA à mesures répétées) : entrez la corrélation entre les mesures (entre les répétitions) dans une ANOVA à mesures répétées.

Correction de la sphéricité (uniquement dans le cas de l'ANOVA à mesures répétées) : entrez la correction appliquée au modèle en cas de non-sphéricité (si l'hypothèse de sphéricité n'est pas rejetée alors on prendra epsilon=1).

Nombre de prédicteurs testés (uniquement dans le cas de l'ANCOVA) : entrez le nombre de variables explicatives quantitatives inclus dans le modèle.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Onglet **Taille de l'effet**:

Recherche de la taille de l'effet : sélectionnez la méthode à utiliser afin de calculer la taille de l'effet (soit en entrant directement la taille de l'effet f, soit à partir des variances du modèle, soit à partir de l'eta² directement, soit en utilisant les moyennes associées à chaque groupe).

Talle de l'effet f (dans le cas où l'on entre directement la taille de l'effet) : entrez la taille de l'effet désirée (voir la partie description pour connaître les conventions).

Variance expliquée (dans le cas où l'on part des variances pour définir la taille de l'effet) : entrez la valeur de la variance expliquée par les variables explicatives étudiées.

Variance de l'erreur (dans le cas où l'on part des variances pour définir la taille de l'effet et uniquement pour l'ANOVA avec interaction et l'ANCOVA) : entrez la valeur de la variance de l'erreur du modèle complet.

Variance intra-groupes (dans le cas où l'on part des variances pour définir la taille de l'effet, dans les autres cas) : entrez la valeur de la variance intra-groupe pour ce modèle.

eta² partiel (dans le cas où l'on utilise la méthode directe afin de calculer la taille de l'effet) : entrez la valeur du eta² partiel (voir la partie description de cette aide).

Ecart-type intra-groupe (dans le cas où l'on part des moyennes et pour l'ANOVA à un facteur et l'analyse inter-facteurs de l'ANOVA à mesures répétées) : entrez la valeur de l'écart-type intra-groupe du modèle.

Les champs suivant apparaîssent lorsqu'on sélectionne comme test, l'ANOVA à un facteur et l'analyse inter-facteurs de l'ANOVA à mesures répétées et comme méthode pour calculer la taille de l'effet, les moyennes.

Moyennes : sélectionnez une colonne de valeur correspondant aux moyennes pour chaque groupe. Ce vecteur doit avoir un nombre de lignes égal au nombre de groupes. Il ne faut pas sélectionner le libellé de la colonne mais uniquement les valeurs.

Tailles des groupes inégales : activez cette option si les groupes ont des tailles différentes. Dans ce cas, il faut sélectionner une colonne de valeurs représentant les tailles de chaque groupe. Cette colonne doit avoir un nombre de lignes égal au nombre de groupes et la somme des éléments doit être égale à la taille de l'échantillon. Cette option n'est pas accessible si l'on recherche la taille de l'échantillon. Il ne faut pas sélectionner le libellé des colonnes ni des lignes mais uniquement les valeurs.

Tailles des groupes égales : activez cette option si les groupes ont des tailles constantes.

Onglet **Graphiques** :

Graphiques de simulation : activez cette option si vous voulez obtenir un graphique en faisant varier différents paramètres du test. On fait varier 2 paramètres, tous les autres paramètres sont utilisés tels qu'ils ont été définis dans l'onglet général.

Axe des X : sélectionnez le paramètre à afficher sur l'axe des abscisses (X). On peut choisir entre la puissance, la taille de l'échantillon, l'erreur de première espèce (alpha) et la taille de l'effet. En fonction de ce que l'on cherche, on aura sur l'axe des Y soit la puissance, soit la taille de l'échantillon.

Taille des intervalles : sélectionnez les bornes inférieures et supérieures de l'axe des X et la taille de l'intervalle entre chaque calcul des paramètres.

Résultats

Entrées : dans ce tableau sont affichées les paramètres ayant permis de calculer la puissance ou le nombre d'observations nécessaires.

Résultats : dans ce tableau sont affichées l'alpha, ta taille de l'effet ainsi que la puissance et le nombre d'observations nécessaires. Les paramètres obtenus par le calcul apparaissent en gras. Une phrase explicative est affichée en dessous de ce tableau.

Intervalles de simulation : ce tableau est composé de 2 colonnes, la puissance, la taille de l'échantillon ou l'alpha en fonction des paramètres sélectionnés dans la boîte de dialogue. Il permet de construire le graphique de simulation.

Graphique de simulation : Ce graphique permet de visualiser l'évolution des paramètres tel que définis dans l'onglet graphiques de la boîte de dialogue.

Exemple

Un exemple de calcul de puissance basé sur un test est disponible sur le Centre d'aide XLSTAT à l'adresse

http://www.xlstat.com/demo-pwrf.htm

Un exemple de calcul de la taille d'échantillon nécessaire est disponible sur le Centre d'aide XLSTAT à l'adresse

http://www.xlstat.com/demo-splf.htm

Bibliographie

Brent R. P (1973). Algorithms for Minimization Without Derivatives. Englewood Cliffs, NJ: Prentice-Hall.

Cohen J. (1988). Statistical Power Analysis for the Behavioral Sciences, Psychology Press, 2-nd Edition.

Sahai H. and Ageel M.I. (2000). The Analysis of Variance. Birkhaüser, Boston.

Régression logistique (Puissance et taille d'échantillon)

Utilisez cet outil pour calculer la puissance ou le nombre d'observations nécessaires lors d'une régression logistique.

Dans cette section :

Description

XLSTAT propose un outil permettant d'appliquer un modèle de régression logistique entre une variable réponse binaire et des variables explicatives quantitatives ou qualitatives. XLSTAT permet également d'estimer la puissance ou de calculer le nombre d'observations nécessaires dans le cadre de cette méthode.

Lorsqu'on teste une hypothèse à l'aide d'un test statistique, on a plusieurs éléments à choisir :

- L'hypothèse nulle H0 et l'hypothèse alternative Ha
- Le test statistique à utiliser
- L'erreur de première espèce (erreur de type I) que l'on appelle aussi $alpha$. Elle se produit lorsqu'on rejette l'hypothèse nulle alors qu'elle est vraie. Elle est fixée a priori pour chaque test et vaut 5%.

L'erreur de seconde espèce ou beta est moins étudiée mais elle revêt une grande importance. En effet, elle représente la probabilité que l'on ne rejette pas l'hypothèse nulle alors qu'elle est fausse. On ne peut pas la fixer a priori mais, on peut essayer de la minimiser, en jouant sur les autres paramètres du modèle. La puissance d'un test est calculée comme 1-beta et représente la probabilité que l'on rejette l'hypothèse nulle alors qu'elle est bien fausse.

On voudra donc maximiser la puissance du test. XLSTAT permet de calculer cette puissance (ainsi que beta) lorsque les autres paramètres du test sont connus. D'autre part, il permet pour une puissance donnée d'évaluer la taille de l'échantillon nécessaire à l'obtention de cette puissance.

Les calculs de puissance en statistique se font généralement avant que l'expérience ne soit menée. On s'en sert principalement pour estimer le nombre d'observations nécessaire pour que l'expérience ait la qualité statistique requise.

Dans le cadre du modèle de régression logistique, la probabilité P de survenue de l'évènement (en général Y=1) est donnée par :

$$P = \frac{\exp(\beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k)}{1 + \exp(\beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k)}$$

On a donc :

$$\log(\frac{P}{1-P}) = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k$$

XLSTAT permet donc de tester si le coefficient β_1 du modèle de régression logistique est égal à 0. Pour plus de détails sur ce modèle, voir le chapitre sur ce sujet.

Nous allons donc tester l'hypothèse :

- H0 : $\beta_1 = 0$
- Ha : $\beta_1 \neq 0$

La puissance du test est calculée à l'aide d'une approximation et dépend du type de variable.

Si X_1 est supposé quantitative et suit une distribution normale, les paramètres utilisés seront :

- P0 (probabilité de référence) : La probabilité que $Y = 1$ lorsque toutes les variables explicatives sont à leur moyenne.
- P1 (probabilité alternative) : La probabilité que X_1 soit égale à une fois son écart-type au-dessus de sa moyenne, sachant que les autres variables explicatives sont à leur moyenne.
- Odds ratio: Le rapport entre la probabilité $Y = 1$, d'une part, lorsque X_1 vaut une fois son écart-type au-dessus de sa moyenne et, d'autre part, lorsque X_1 est à sa moyenne.
- Le R² obtenu en faisant une régression entre X_1 et les autres variables explicatives du modèle.

Si X_1 est binaire et suit une loi binomiale. Les paramètres utilisés seront :

- P0 (probabilité de référence) : La probabilité que $Y = 1$ sachant que $X_1 = 0$.
- P1 (probabilité alternative) : La probabilité que $Y = 1$ sachant que $X_1 = 1$.
- Odds ratio: Le rapport entre la probabilité $Y = 1$ lorsque $X_1 = 1$ et lorsque X_1=0.
- Le R² obtenu en faisant une régression entre X_1 et les autres variables explicatives du modèle.
- Le pourcentage d'observations telles que $X_1 = 1$.

Ces approximations dépendent de la loi normale et de ces paramètres et permettent de calculer la puissance de ce test.

Calcul de la taille de l'échantillon

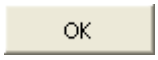
Afin de calculer le nombre d'observations nécessaires, XLSTAT utilise un algorithme de recherche de racine d'une fonction appelé algorithme Van Wijngaarden-Dekker-Brent (Brent, 1973). Cet algorithme est adapté au cas où les dérivées de la fonction ne sont pas connues. On cherche ainsi N tel que

puissance_test(N)-puissance_recherchée=0

On obtient donc la taille N telle que la puissance soit la plus proche possible de la puissance recherchée.

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.

Onglet **Général**:

Objectif : choisissez entre le calcul de la puissance et la recherche du nombre d'observations (en fonction de ce choix les champs suivants seront différents).

Alpha : entrez l'erreur de première espèce (elle doit être comprise entre 0,001 et 0,999).

Puissance (dans le cas où l'on recherche la taille de l'échantillon) : entrez la puissance requise (elle doit être comprise entre 0,001 et 0,999).

Taille d'échantillon (dans le cas où l'on cherche à calculer la puissance) : entrez la taille de l'échantillon.

Probabilité de référence (P0) : entrez la probabilité de référence, c'est à-dire la probabilité que $Y = 1$ lorsque toutes les variables explicatives sont à leur moyenne ou à 0 si celles-ci sont binaires.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Onglet **Taille de l'effet**:

Recherche de la taille de l'effet : sélectionnez la méthode à utiliser afin de calculer la taille de l'effet (soit en entrant la probabilité alternative P1, soit à partir de l'odds ratio).

Probabilité alternative (P1) : entrez la probabilité que X_1 soit égal à une fois son écart-type au-dessus de sa moyenne si X_1 est quantitative ou que $X_1 = 1$ si X_1 est binaire.

Odds ratio : entrez la valeur de l'odds ratio (voir la partie description pour plus de détails).

R² de X_1 avec les autres X : entrez la valeur du R² obtenu par la régression entre X_1 et les autres variables explicatives du modèle de Cox.

Type de variable : sélectionnez le type de variable analysée, quantitative ou binaire.

Pourcentage de N avec X_1=1 (uniquement dans le cas d'une variable binaire): entrez le pourcentage (entre 1 et 99) d'observations pour lesquelles X_1=1.

Onglet **Graphiques** :

Graphiques de simulation : activez cette option si vous voulez obtenir un graphique en faisant varier différents paramètres du test. On fait varier 2 paramètres, tous les autres paramètres sont utilisés tels qu'ils ont été définis dans l'onglet général.

Axe des X : sélectionnez le paramètre à afficher sur l'axe des abscisses (X). On peut choisir entre la puissance, la taille de l'échantillon, l'erreur de première espèce (alpha) et la taille de l'effet. En fonction de ce que l'on cherche, on aura sur l'axe des Y soit la puissance, soit la taille de l'échantillon.

Taille des intervalles : sélectionnez les bornes inférieures et supérieures de l'axe des X et la taille de l'intervalle entre chaque calcul des paramètres.

Résultats

Entrées : dans ce tableau sont affichées les paramètres ayant permis de calculer la puissance ou le nombre d'observations nécessaires.

Résultats : dans ce tableau sont affichées l'alpha ainsi que la puissance et le nombre d'observations nécessaires. Les paramètres obtenus par le calcul apparaissent en gras. Une phrase explicative est affichée en dessous de ce tableau.

Intervalles de simulation : ce tableau est composé de 2 colonnes, la puissance, la taille de l'échantillon ou l'alpha en fonction des paramètres sélectionnés dans la boîte de dialogue. Il permet de construire le graphique de simulation.

Graphique de simulation : Ce graphique permet de visualiser l'évolution des paramètres tel que définis dans l'onglet graphiques de la boîte de dialogue.

Exemple

Un exemple de calcul de puissance basé sur un test est disponible sur le Centre d'aide XLSTAT à l'adresse

http://www.xlstat.com/demo-pwrf.htm

Un exemple de calcul de la taille d'échantillon nécessaire est disponible sur le Centre d'aide XLSTAT à l'adresse

http://www.xlstat.com/demo-splf.htm

Bibliographie

Brent R. P (1973). Algorithms for Minimization Without Derivatives. Englewood Cliffs, NJ: Prentice-Hall.

Cohen J. (1988). Statistical Power Analysis for the Behavioral Sciences, Psychology Press, 2-nd Edition.

Hosmer D.W. and Lemeshow S. (2000). Applied Logistic Regression, Second Edition. John Wiley and Sons, New York.

Modèle de Cox (Puissance et taille d'échantillon)

Utilisez cet outil pour calculer la puissance ou le nombre d'observations nécessaires lors de l'application du modèle de Cox en analyse de données de survie.

Dans cette section :

Description

XLSTAT propose un outil permettant d'appliquer le modèle à risques proportionnels de Cox sur des données de survie. XLSTAT permet également d'estimer la puissance ou de calculer le nombre d'observations nécessaires dans le cadre de cette approche.

Lorsqu'on teste une hypothèse à l'aide d'un test statistique, on a plusieurs éléments à choisir :

- L'hypothèse nulle H0 et l'hypothèse alternative Ha
- Le test statistique à utiliser
- L'erreur de première espèce (erreur de type I) que l'on appelle aussi $alpha$. Elle se produit lorsqu'on rejette l'hypothèse nulle alors qu'elle est vraie. Elle est fixée a priori pour chaque test et vaut 5%.

L'erreur de seconde espèce ou beta est moins étudiée mais elle revêt une grande importance. En effet, elle représente la probabilité que l'on ne rejette pas l'hypothèse nulle alors qu'elle est fausse. On ne peut pas la fixer a priori mais, on peut essayer de la minimiser, en jouant sur les autres paramètres du modèle. La puissance d'un test est calculée comme 1-beta et représente la probabilité que l'on rejette l'hypothèse nulle alors qu'elle est bien fausse.

On voudra donc maximiser la puissance du test. XLSTAT permet de calculer cette puissance (ainsi que beta) lorsque les autres paramètres du test sont connus. D'autre part, il permet pour une puissance donnée d'évaluer la taille de l'échantillon nécessaire à l'obtention de cette puissance.

Les calculs de puissance en statistique se font généralement avant que l'expérience ne soit menée. On s'en sert principalement pour estimer le nombre d'observations nécessaire pour que l'expérience ait la qualité statistique requise.

Le modèle de Cox exprime le risque instantané de survenue de l'événement $\lambda(t, X_1, X_2, ..., X_p)$ sous la forme :

$$\lambda(t, X) = \lambda_0(t)exp(\beta_1 X_1 + \ldots + \beta_p X_p))$$

Cette formule appelle quelques commentaires. Le risque instantané se décompose en 2 termes dont l'un dépend du temps et l'autre des variables X_j. Si par exemple, les variables X_j représentent des facteurs de risque et si elles sont toutes égales à 0, $\lambda_0(t)$ est le risque instantané de sujets ne présentant aucun facteur de risque. La forme de $\lambda_0(t)$ n'étant pas précisée, c'est plutôt l'association entre les variables X_j et la survenue de l'événement considéré qui est l'intérêt central du modèle. Cela revient à déterminer les coefficients β_j.

XLSTAT permet donc de tester si le coefficient β_1 du modèle de Cox est différent de 0. Si $\beta_1 \neq 0$, alors on pourra dire que le facteur testé est un facteur de risque. Pour plus de détails sur ce modèle, voir le chapitre sur ce sujet.

Nous allons donc tester l'hypothèse :

- H0 : $\beta_1 = 0$
- Ha : $\beta_1 \neq 0$

Pour cela, nous utiliserons la statistique de Wald :

$$z = \frac{\beta_1}{\sqrt{\mathrm{var}(\beta_1)}}$$

La puissance du test est calculée à l'aide d'une approximation et dépend de la loi normale, de la proportion d'individus qui ne sont pas censurés, de la variance de la variable X_1, de la valeur supposée de β_1 notée B et du R^2 obtenu par la régression des autres variables explicatives du modèle sur la variable X_1.

On obtient alors la puissance de ce test.

Calcul de la taille de l'échantillon

Afin de calculer le nombre d'observations nécessaires, XLSTAT utilise un algorithme de recherche de racine d'une fonction appelé algorithme Van Wijngaarden-Dekker-Brent (Brent, 1973). Cet algorithme est adapté au cas où les dérivées de la fonction ne sont pas connues. On cherche ainsi N tel que

puissance_test(N)-puissance_recherchée=0

On obtient donc la taille N telle que la puissance soit la plus proche possible de la puissance recherchée.

Calcul de B

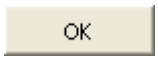
Le coefficient B(Log(Rapport de risque)) est l'estimation du coefficient β_1 tel qu'il apparaît dans l'équation suivante :

$$\log(\frac{\lambda(t|X)}{\lambda_0(t)}) = \beta_1 X_1 + \ldots + \beta_p X_p$$

β_1 représente le changement du logarithme du rapport de risque lorsque X_1 augmente de 1 (toutes les autres variables restant constantes). Si on préfère partir du rapport de risque on prendra une valeur que l'on transformera en son logarithme. Ainsi pour un rapport de risque de 2, on aura $B = ln(2) = 0,693$.

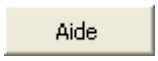
Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.



: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.

Onglet **Général**:

Objectif : choisissez entre le calcul de la puissance et la recherche du nombre d'observations (en fonction de ce choix les champs suivants seront différents).

Alpha : entrez l'erreur de première espèce (elle doit être comprise entre 0,001 et 0,999).

Puissance (dans le cas où l'on recherche la taille de l'échantillon) : entrez la puissance requise (elle doit être comprise entre 0,001 et 0,999).

Taille d'échantillon (dans le cas où l'on cherche à calculer la puissance) : entrez la taille de l'échantillon.

Taux d'évènement : entrez la proportion d'individus non censurés (entre 0,001 et 0,999).

B(log(Rapport de risque) : entrez le paramètre B associé à la variable X_1 dans le modèle de Cox.

Ecart-type de X_1 : entrez l'écart-type de X_1.

R² de X_1 avec les autres X : entrez la valeur du R^2 obtenu par la régression entre X_1 et les autres variables explicatives du modèle de Cox.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Onglet **Graphiques** :

Graphiques de simulation : activez cette option si vous voulez obtenir un graphique en faisant varier différents paramètres du test. On fait varier 2 paramètres, tous les autres paramètres sont utilisés tels qu'ils ont été définis dans l'onglet général.

Axe des X : sélectionnez le paramètre à afficher sur l'axe des abscisses (X). On peut choisir entre la puissance, la taille de l'échantillon, l'erreur de première espèce (alpha) et la taille de l'effet. En fonction de ce que l'on cherche, on aura sur l'axe des Y soit la puissance, soit la taille de l'échantillon.

Taille des intervalles : sélectionnez les bornes inférieures et supérieures de l'axe des X et la taille de l'intervalle entre chaque calcul des paramètres.

Résultats

Entrées : dans ce tableau sont affichées les paramètres ayant permis de calculer la puissance ou le nombre d'observations nécessaires.

Résultats : dans ce tableau sont affichées les paramètres du test ainsi que la puissance ou le nombre d'observations nécessaires. Les paramètres obtenus par le calcul apparaissent en gras. Une phrase explicative est affichée en dessous de ce tableau.

Intervalles de simulation : ce tableau est composé de 2 colonnes, la puissance, la taille de l'échantillon ou l'alpha en fonction des paramètres sélectionnés dans la boîte de dialogue. Il permet de construire le graphique de simulation.

Graphique de simulation : Ce graphique permet de visualiser l'évolution des paramètres tel que définis dans l'onglet graphiques de la boîte de dialogue.

Exemple

Un exemple de calcul de puissance basé sur un test est disponible sur le Centre d'aide XLSTAT à l'adresse

http://www.xlstat.com/demo-pwrf.htm

Un exemple de calcul de la taille d'échantillon nécessaire est disponible sur le Centre d'aide XLSTAT à l'adresse

http://www.xlstat.com/demo-splf.htm

Bibliographie

Brent, R. P (1973). Algorithms for Minimization Without Derivatives. Englewood Cliffs, NJ: Prentice-Hall.

Cohen, J. (1988). Statistical Power Analysis for the Behavioral Sciences, Psychology Press, 2-nd Edition.

Cox D. R. and Oakes D. (1984). Analysis of Survival Data. Chapman and Hall, London.

Kalbfleisch J. D. and Prentice R. L. (2002). The Statistical Analysis of Failure Time Data. 2-nd edition, John Wiley & Sons, New York.

Taille d'échantillon pour les essais cliniques (Puissance et taille d'échantillon)

Utilisez cet outil pour calculer le nombre d'observations nécessaires ou la puissance obtenue pour des essais cliniques du type : test d'équivalence, test de supériorité et test de non infériorité.

Dans cette section :

Description

XLSTAT permet de calculer la taille de l'échantillon pour des tests cliniques lorsqu'on possède certaines informations sur les objectifs et les résultats escomptés.

XLSTAT permet de connaître la taille de l'échantillon pour 3 types d'essais :

- Les tests d'équivalence : lorsqu'on cherche à voir si, par exemple, un nouveau traitement peut en remplacer un ancien.
- Les tests de supériorité : lorsqu'on cherche à voir si, par exemple, un traitement est plus efficace qu'un autre.
- Les tests de non infériorité : lorsqu'on cherche à voir si, par exemple, un nouveau traitement est au moins aussi efficace qu'un ancien traitement avec une marge donnée.

Ces tests peuvent être appliqués soit à une variable binaire, soit à une variable continue.

Lorsqu'on teste une hypothèse à l'aide d'un test statistique, on a plusieurs éléments à choisir :

- L'hypothèse nulle H0 et l'hypothèse alternative Ha
- Le test statistique à utiliser
- L'erreur de première espèce (erreur de type I) que l'on appelle aussi $alpha$. Elle se produit lorsqu'on rejette l'hypothèse nulle alors qu'elle est vraie. Elle est fixée a priori pour chaque test et vaut 5%.

L'erreur de seconde espèce ou beta est moins étudiée mais elle revêt une grande importance. En effet, elle représente la probabilité que l'on ne rejette pas l'hypothèse nulle alors qu'elle est fausse. On ne peut pas la fixer a priori mais, on peut essayer de la minimiser, en jouant sur les

autres paramètres du modèle. La puissance d'un test est calculée comme 1-beta et représente la probabilité que l'on rejette l'hypothèse nulle alors qu'elle est bien fausse.

On voudra donc maximiser la puissance du test. XLSTAT permet de calculer cette puissance (ainsi que beta) lorsque les autres paramètres du test sont connus. D'autre part, il permet pour une puissance donnée d'évaluer la taille de l'échantillon nécessaire à l'obtention de cette puissance.

Les calculs de puissance en statistique se font généralement avant que l'expérience ne soit menée. On s'en sert principalement pour estimer le nombre d'observations nécessaire pour que l'expérience ait la qualité statistique requise.

Méthodes

La taille de l'échantillon nécessaire à l'application de ces tests est obtenue en utilisant des formules simples.

Test d'équivalence pour une variable continue

On compare les moyennes associées à une variable continue entre deux groupes randomisés. L'hypothèse nulle est telle que les traitements sont équivalents avec une marge d définie par l'utilisateur.

Pour obtenir la taille de l'échantillon, on utilise la formule suivante :

$$n = \frac{f(\alpha, \beta/2) \cdot 2 \cdot \sigma^2}{(d)^2}$$

Avec σ^2 variance pour les deux groupes. On a :

$$f(\alpha, \beta) = (\Phi^{-1}(\alpha) + \Phi^{-1}(\beta))^2$$

Test d'équivalence pour une variable binaire

On compare le pourcentage de « survie » entre deux groupes randomisés. L'hypothèse nulle est que les traitements sont équivalents avec une marge d définie par l'utilisateur.

Pour obtenir la taille de l'échantillon, on utilise la formule suivante :

$$n = \frac{f(\alpha, \beta/2) \cdot (P(std) \cdot (100 - P(std)))}{(P(std) - d)^2}$$

Avec $P(std)$, pourcentage de « survie » pour les deux traitements étudiés. d est la marge acceptée pour la différence entre les deux traitements.

On a :

$$f(\alpha, \beta) = (\Phi^{-1}(\alpha) + \Phi^{-1}(\beta))^2$$

Test de non infériorité pour une variable continue

On compare les moyennes associées à une variable continue entre deux groupes randomisés. L'hypothèse nulle est telle que la moyenne avec le traitement de base est meilleure avec une marge de d comparé à la moyenne associée au nouveau traitement.

L'hypothèse alternative est équivalente à dire que le nouveau traitement est meilleur que le traitement de base ou un tout petit peu moins bon d'une valeur de d. Cette valeur doit être fixée en fonction de la situation.

Pour obtenir la taille de l'échantillon, on utilise la formule suivante :

$$n = \frac{f(\alpha, \beta) \cdot 2 \cdot \sigma^2}{(d)^2}$$

Avec σ^2 variance pour les deux groupes. On a :

$$f(\alpha, \beta) = (\Phi^{-1}(\alpha) + \Phi^{-1}(\beta))^2$$

Test de non infériorité pour une variable binaire

On compare le pourcentage de « survie » entre deux groupes randomisés. L'hypothèse nulle est telle que le pourcentage de survie avec le traitement de base est meilleure avec une marge de d comparé au pourcentage associé au nouveau traitement.

L'hypothèse alternative est équivalente à dire que le nouveau traitement est meilleur que le traitement de base ou un tout petit peu moins bon d'une valeur de d. Cette valeur doit être fixée en fonction de la situation.

Pour obtenir la taille de l'échantillon, on utilise la formule suivante :

$$n = \frac{f(\alpha, \beta).P(std)(100 - P(Std)) + P(new)(100 - P(new))}{(P(Std) - P(new) - d)^2}$$

Avec $P(std)$, pourcentage de « survie » pour le traitement de base (standard) et $P(new)$, pourcentage de « survie » pour le nouveau traitement. d est le seuil d'infériorité.

On a :

$$f(\alpha, \beta) = (\Phi^{-1}(\alpha) + \Phi^{-1}(\beta))^2$$

Test de supériorité pour une variable continue

On compare la moyenne de deux groupes randomisés. On teste le fait de savoir si la moyenne associée à une variable continue est plus grande pour le nouveau traitement par rapport à celle associée à l'ancien traitement.

Pour obtenir la taille de l'échantillon, on utilise la formule suivante :

$$n = \frac{f(\alpha/2, \beta) \cdot 2 \cdot \sigma^2}{(\mu_1 - \mu_2)^2}$$

Avec σ^2, variance des groupes et mu1, mu2 moyennes des deux groupes. On a :

$$f(\alpha, \beta) = (\Phi^{-1}(\alpha) + \Phi^{-1}(\beta))^2$$

Dans le cas d'existence de cross-over, on utilise une formule pour l'ajustement de la taille de l'échantillon :

$$n_{\text{adjusted}} = \frac{n^* 10000}{(100 - c_1 - c_2)}$$

Avec c1 et c2 pourcentage de cross-over dans chacun des groupes de patients.

Test de supériorité pour une variable binaire

On compare le pourcentage de « survie » entre deux groupes randomisés. L'hypothèse nulle est telle que le pourcentage de survie avec le traitement de base est meilleure comparé au pourcentage associé au nouveau traitement.

Pour obtenir la taille de l'échantillon, on utilise la formule suivante :

$$n = \frac{f(\alpha/2, \beta) \cdot (P(std) \cdot (100 - P(std)) + P(new) \cdot (100 - P(new)))}{(P(std) - P(new))^2}$$

Avec $P(std)$, pourcentage de « survie » pour le traitement de base (standard) et $P(new)$, pourcentage de « survie » pour le nouveau traitement.

On a :

$$f(\alpha, \beta) = (\Phi^{-1}(\alpha) + \Phi^{-1}(\beta))^2$$

Dans le cas d'existence de cross-over, on utilise une formule pour l'ajustement de la taille de l'échantillon :

$$n_{\text{adjusted}} = \frac{n^* 10000}{(100 - c_1 - c_2)}$$

Avec c1 et c2 pourcentage de cross-over dans chacun des groupes de patients.

Calcul de la puissance

Afin de calculer la puissance, XLSTAT utilise un algorithme de recherche simple. On cherche ainsi beta tel que

Taille d'échantillon(beta)-taille d'échantillon recherchée=0

On obtient donc la puissance telle que la taille d'échantillon soit la plus proche possible de la taille de l'échantillon recherchée.

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.

OK : cliquez sur ce bouton pour lancer les calculs.

Annuler : cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

Aide : cliquez sur ce bouton pour afficher l'aide.

 : cliquez sur ce bouton pour rétablir les options par défaut.

 : cliquez sur ce bouton pour effacer les sélections de données.

Onglet **Général**:

Objectif : choisissez entre le calcul de la puissance et la recherche du nombre d'observations (en fonction de ce choix les champs suivants seront différents).

Test clinique : sélectionnez le test clinique à utiliser. Equivalence, supériorité ou non infériorité.

Variable à expliquer : choisissez le type de variable à expliquer (continue ou binaire).

Alpha : entrez l'erreur de première espèce (elle doit être comprise entre 0,001 et 0,999).

Taille de l'essai (dans le cas où l'on cherche à calculer la puissance) : entrez la taille de l'essai, c'est-à-dire le nombre total de sujets analysés dans l'essai clinique.

Puissance (dans le cas où l'on recherche la taille de l'échantillon) : entrez la puissance requise (elle doit être comprise entre 0,001 et 0,999).

Les options suivantes dépendent de l'essai effectué ainsi que du type de variable dépendante.

Test d'équivalence pour une variable continue

Ecart-type: entrez l'écart type de la variable continue.

Limite d'équivalence : entrez la valeur d pour définir l'équivalence entre les groupes.

Test d'équivalence pour une variable binaire

% de survie pour les groupes: entrez le pourcentage de réussite (survie) pour les deux groupes de patients. On suppose qu'il est égal.

Limite d'équivalence : entrez la valeur d pour définir l'équivalence entre les groupes.

Test de non infériorité pour une variable continue

Ecart-type: entrez l'écart type de la variable continue.

Limite de non infériorité : entrez la valeur d pour définir la limite de non infériorité entre les deux groupes.

Test de non infériorité pour une variable binaire

% de survie pour le groupe contrôle: entrez le pourcentage de réussite (survie) pour le groupe contrôle.

% de survie pour le groupe traité: entrez le pourcentage de réussite (survie) pour le groupe traité.

Limite de non infériorité : entrez la valeur d pour définir la limite de non infériorité entre les deux groupes.

Test de supériorité pour une variable continue

Moyenne pour le groupe contrôle : entrez la moyenne au niveau du groupe contrôle.

Moyenne pour le groupe traité : entrez la moyenne au niveau du groupe traité.

Ecart-type: entrez l'écart type de la variable continue.

% de cross-over pour le groupe contrôle: entrez le pourcentage de cross- over pour le groupe contrôle.

% de cross-over pour le groupe traité: entrez le pourcentage de cross- over pour le groupe traité.

Test de supériorité pour une variable binaire

% de survie pour le groupe contrôle: entrez le pourcentage de réussite (survie) pour le groupe contrôle.

% de survie pour le groupe traité: entrez le pourcentage de réussite (survie) pour le groupe traité.

% de cross-over pour le groupe contrôle: entrez le pourcentage de cross- over pour le groupe contrôle.

% de cross-over pour le groupe traité: entrez le pourcentage de cross- over pour le groupe traité.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Onglet Graphiques :

Graphiques de simulation : activez cette option si vous voulez obtenir un graphique en faisant varier différents paramètres du test. On fait varier 2 paramètres, tous les autres paramètres sont utilisés tels qu'ils ont été définis dans l'onglet général.

Axe des X : sélectionnez le paramètre à afficher sur l'axe des abscisses (X). On peut choisir entre la puissance, la taille de l'échantillon, l'erreur de première espèce (alpha) et la taille de l'effet. En fonction de ce que l'on cherche, on aura sur l'axe des Y soit la puissance, soit la taille de l'échantillon.

Taille des intervalles : sélectionnez les bornes inférieures et supérieures de l'axe des X et la taille de l'intervalle entre chaque calcul des paramètres.

Résultats

Résultats : dans ce tableau sont affichés les paramètres du test ainsi que la puissance ou le nombre d'observations nécessaires. Les paramètres obtenus par le calcul apparaissent en gras. Une phrase explicative est affichée en dessous de ce tableau.

Intervalles de simulation : ce tableau est composé de 2 colonnes, la puissance, la taille de l'échantillon ou l'alpha en fonction des paramètres sélectionnés dans la boîte de dialogue. Il permet de construire le graphique de simulation.

Graphique de simulation : Ce graphique permet de visualiser l'évolution des paramètres el que définis dans l'onglet graphiques de la boîte de dialogue.

Exemple

Un exemple de calcul de la taille d'échantillon nécessaire pour des essais cliniques est disponible sur le Centre d'aide XLSTAT à l'adresse

http://www.xlstat.com/demo-splctrialf.htm

Bibliographie

Blackwelder W.C. (1982). Providing the null hypothesis in Clinical trials. *Control. Clin. Trials*, **3**, 345-353.

Cohen J. (1988). Statistical Power Analysis for the Behavioral Sciences, Psychology Press, 2-nd Edition.

Pocock S.J. (1983). Clinical trials : a practical approach, Wiley.

Maîtrise Statistique des Procédés

Cartes pour sous-groupes

Utilisez cet outil pour maîtriser la qualité de vos procédés dans le cas où vous avez plusieurs mesures pour chaque pas de temps. Les mesures doivent être des données numériques. Cette méthode est utile pour résumer l'évolution de la moyenne et de la variabilité de la qualité d'une production.

Vous trouverez intégrés à cet outil, les transformations Box-Cox et le calcul de capacité du processus, ainsi que la possibilité d'appliquer des règles spéciales ou des règles de Westgard (un ensemble de règles alternatives pour identifier des causes spéciales) pour compléter votre analyse.

Dans cette section :

Description

Les cartes de contrôle ont été d'abord mentionnées par Walter Shewhart dans un document écrit alors qu'il travaillait aux Bell Labs en 1924. Il a ensuite décrit ses méthodes plus en détails dans un livre (1931).

Pendant plusieurs années, il n'y eu pas d'avancées majeures dans ce domaine, jusqu'à ce que Deming mette au point les cartes de contrôle CUSUM, UWMA et EWMA en 1936.

Les cartes de contrôle étaient à l'origine utilisées pour le contrôle de la qualité des biens de production. Pour cette raison, le vocabulaire utilisé pour ces méthodes statistiques provient souvent de ce domaine d'application. Aujourd'hui, ces approches sont appliquées dans de nombreux autres domaines, comme par exemple les services, les ressources humaines ou les ventes. Dans les lignes qui suivent nous utilisons le vocabulaire du domaine de la production.

Cartes pour sous-groupes

L'outil de création de cartes de contrôle pour sous-groupes permet de créer les graphiques suivants, seuls ou combinés :

- $\overline{X}$ (X barre) : la carte X barre permet de suivre l'évolution de la moyenne d'un procédé de production. Des décalages de la moyenne sont aisément visibles sur de tels graphiques.
- R : la carte R (Range chart en anglais) est utile pour analyser la variabilité d'une production. Une variation importante de la qualité de la production, provoquée par exemple par l'utilisation de différentes chaînes de production sera facilement détectable.
- S, S² : les cartes S et S² sont aussi utilisées pour contrôler la variabilité de la production. Sur la carte de contrôle S on représente l'écart-type du processus suivi, tandis que sur la carte S² on suit la variance (le carré de l'écart-type).

Remarque 1 : si vous souhaitez pouvoir détecter des décalages plus faibles de la moyenne, vous pouvez utiliser les cartes CUSUM qui sont d'ailleurs souvent préférées aux cartes pour sous-groupes.

Remarque 2 : si vous ne disposez que d'une seule mesure pour chaque pas de temps, vous devez utiliser les cartes de contrôle à valeurs individuelles.

Remarque 3 : si vos mesures sont de nature qualitative (par exemple, oui/non, conforme/non conforme), vous devez utiliser les cartes de contrôle par attributs.

XLSTAT vous propose les options suivantes pour l'estimation de l'écart-type (sigma ou σ) d'un échantillon, pour k sous-groupes et $n_i (i = 1, \ldots, k)$ mesures par sous-groupe :

- Ecart-type global : sigma est calculé à partir des k variances intra-sous-groupes, selon la formule suivante :

$$\hat{s} = \sqrt{\frac{\sum_{i=1}^{k}(n_i - 1)s_i^2}{\sum_{i=1}^{k}(n_i - 1)}} / c_4 \left(1 + \sum_{i=1}^{k}(n_i - 1)\right)$$

où c_4 est une constante définie par Burr (1969).

- R-barre : l'estimateur de sigma est calculé à partir de l'étendue moyenne (ou amplitude moyenne) pour les k sous-groupes.

$$\hat{s} = \overline{R}/d_2$$

où d_2 est une constante définie par Burr (1969).

- S-barre : sigma est calculé à partir de la moyenne des k variances intra-sous-groupes, selon la formule suivante :

$$\hat{s} = \sqrt{\frac{1}{k}\sum_{i=1}^{k} s_i^2} / c_4$$

où c_4 est une constante définie par Burr (1969).

Capabilité du procédé

La capabilité du procédé décrit un procédé (ou processus) et permet de savoir s'il est sous contrôle et si les données correspondant aux variables mesurées sont à l'intérieur des limites de spécification du procédé. Dans un tel cas, on dit que le procédé est « capable ».

Dans un contexte industriel, la capabilité d'un procédé est sa capacité à produire des pièces dont les caractéristiques sont à l'intérieur d'un intervalle de tolérance. La capabilité du procédé permet de comparer la dispersion du procédé à l'intervalle de tolérance.

Au cours de l'interprétation des différents indices de capabilité des procédés, veuillez prendre garde au fait que certains indicateurs nécessitent de faire l'hypothèse de normalité ou, tout au moins, de la symétrie de la distribution des variables mesurées. En utilisant les tests de normalité vous pourrez vérifier la validité de ces hypothèses (voir les Tests de Normalité).

Si l'hypothèse de normalité ne peut être retenue, vous avez les possibilités suivantes pour obtenir des capabilités des procédés :

- Utiliser une transformation Box-Cox pour améliorer la normalité des échantillons, et vérifier ensuite à nouveau la normalité avec un test.
- Utiliser l'indicateur de capabilité de procédé Cp 5.15.

Soit m, s, LSL, USL respectivement les estimateurs de la moyenne, l'écart-type, la limite de spécification inférieure, la limite de spécification supérieure du procédé, et T la cible choisie. Soit μ et σ les moyenne et écart-type théoriques du procédé. XLSTAT permet de calculer les indices de performance suivants pour évaluer la capabilité du procédé :

- Cp : L'indice de capabilité court terme du procédé est estimé par :

$$\hat{C}p = \frac{USL - LSL}{6s}$$

- Cpl : L'indice de capabilité court terme inférieure du procédé est estimé par :

$$\hat{C}pl = \frac{m - LSL}{3s}$$

- Cpu : L'indice de capabilité court terme supérieure du procédé est estimé par :

$$\hat{C}pu = \frac{USL - m}{3s}$$

- Cpk : Cet indice de capabilité court terme, qui contrairement au Cp nécessite la connaissance de la moyenne, est estimé par :

$$\hat{C}pk = min(Cpl, Cpu)$$

- Pp : La capabilité long terme du procédé est définie par:

$$Pp = \frac{USL - LSL}{6\sigma}$$

- Ppl : La capabilité long terme inférieure du procédé est définie par:

$$Ppl = \frac{\mu - LSL}{3\sigma}$$

- Ppu : La capabilité long terme supérieure du procédé est définie par:

$$Ppu = \frac{USL - \mu}{3\sigma}$$

- Ppk : La capabilité long terme Ppk, qui contrairement au Pp nécessite la connaissance de la moyenne, est défini par :

$$Ppk = min(Ppl, Ppu)$$

- Cpm : L'indice de capabilité court terme de Taguchi peut être calculé si une valeur cible (T) a été spécifiée. Son estimateur est défini par :

$$\hat{C}pm = \frac{USL - LSL}{6\sqrt{s^2 + (m - T)^2}}$$

- $CpmBoyles$: L'indice de capabilité court terme de Taguchi amélioré par Boyles (1991) qui nécessite également la spécification d'une valeur cible (T), a pour estimateur :

$$\hat{C}pm\ Boyles = \frac{USL - LSL}{6\sqrt{(n-1)s^2/n + (m - T)^2}}$$

- $Cp5.15$: Cet indice de capabilité court terme est défini par :

$$\hat{C}p\ 5.15 = \frac{USL - LSL}{5.15s}$$

- $Cpk5.15$: Cet indice de capabilité court terme est défini par :

$$\hat{C}pk\ 5.15 = \frac{(USL - LSL)/2 - |m - (USL + LSL)/2|}{2.57s}$$

- $Cpmk$: Cet indice de capabilité court terme a été proposé par Pearn. Il peut être calculé si la valeur cible a été spécifiée :

$$\hat{C}pmk = \frac{s \cdot Cpk}{\sqrt{s^2 + (m - T)^2}}$$

- $CsWright$: La capabilité du procédé telle que proposée par Wright (1995). Elle peut être calculée si la valeur cible a été spécifiée :

$$\hat{C}s = \frac{(USL - LSL)/2 - |m - T|}{3\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - T)^2 + \left|\frac{n^2 m_3}{(n-1)(n-2)} \times \left(\frac{n}{n-1} \times \frac{m_2}{c_4^2}\right)^{-1/2}\right|}}$$

avec

$$m_r = \sum_{i=1}^{n}(x_i - T)^r$$

$$c_4 = \left[\frac{2}{n-1}\right]^{1/2} \Gamma(\frac{n}{2})/\Gamma(\frac{n-1}{2})^{-1}$$

- Z below: Le nombre d'écarts-types entre la moyenne et la limite de spécification inférieure. Il est défini par :

$$Zbelow = (m - LSL)/s$$

- Z above: Le nombre d'écarts-types entre la moyenne et la limite de spécification supérieure. Il est défini par :

$$Zabove = (USL - m)/s$$

- Z total: Le nombre d'écarts-types entre les limites de spécification inférieure et supérieure. Il est défini par:

$$Ztotal = (USL - LSL)/s$$

- p(not conform) below: La probabilité d'avoir un produit défectueux sous la limite de spécification inférieure :

$$p(\text{not conform})below = \Phi(Zbelow)$$

- p(not conform) above: La probabilité d'avoir un produit défectueux au-delà de la limite de spécification supérieure :

$$p(\text{not conform})above = \Phi(Zabove)$$

- p(not conform) total: La probabilité d'avoir un produit défectueux sous la limite de spécification inférieure ou au-delà de la limite de spécification supérieure :

$$p(\text{not conform})total = p(\text{not conform})below + p(\text{not conform})above$$

- PPM below: Le nombre de produits défectueux sous la limite de spécification inférieure pour une production d'un million de produits :

$$PPMbelow = p(notconform)below \times 10^6$$

- PPM above: Le nombre de produits défectueux au-delà de la limite de spécification supérieure pour une production d'un million de produits :

$$PPMabove = p(notconform)above \times 10^6$$

- PPM total: Le nombre de produits défectueux hors des limites de spécification pour une production d'un million de produits :

$$PPMtotal = PPMbelow + PPMabove$$

Transformation Box-Cox

Transformation Box-Cox : activez cette option pour faire une transformation de Box-Cox. Vous pouvez soit imposer une valeur de λ, soit décider que XLSTAT doit l'optimiser. Cette transformation permet d'augmenter la normalité des données ; l'équation de Box-Cox est définie par :

$$Y_t = \begin{cases} \frac{X_t^\lambda - 1}{\lambda} & (X_t > 0, \lambda \neq 0) \text{ ou } (X_t \geq 0, \lambda > 0) \\ ln(X_t) & X_t > 0, \lambda = 0 \end{cases}$$

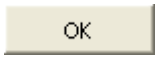
Si l'option d'optimisation est choisie, XLSTAT détermine la valeur de λ qui maximise la vraisemblance de l'échantillon, étant supposé qu'après transformation l'échantillon suit une loi normale.

Règles pour l'interprétation des cartes

XLSTAT vous donne la possibilité d'appliquer des règles pour les « causes spéciales » ainsi que les règles de Westgard. Deux ensembles de règles sont proposés pour l'interprétation des graphiques. Vous pouvez activer ou désactiver les règles pour chacun d'eux.

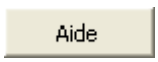
Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.



: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les

variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.

 : cliquez sur ce bouton pour changer la façon dont vous chargez les données dans XLSTAT. Si le bouton possède une icône de souris, XLSTAT vous permet de faire une sélection à la souris des données. Si le bouton possède une icône de liste, XLSTAT vous permet de sélectionner les données à partir d'une liste. Si le bouton possède une icône orange, des boutons complémentaires avec un point d'interrogation  sont affichés pour vous permettre d'importer les données à partir de fichiers.

Onglet **Général** :

Type de carte : choisissez le type de carte de contrôle que vous voulez utiliser :

- **Carte X barre** : choisissez cette option si vous voulez calculer une carte X barre pour analyser la moyenne d'un procédé.
- **Carte R chart** : choisissez cette option si vous voulez calculer une carte X pour analyser la variabilité d'un processus.
- **Carte S chart** : choisissez procédé option si vous voulez calculer une carte S pour analyser la variabilité d'un procédé.
- **Carte S² chart** : choisissez cette option si vous voulez calculer une carte S² pour analyser la variabilité d'un procédé.
- **Carte X barre R chart** : choisissez cette option si vous voulez calculer une carte X barre combinée avec une carte R pour analyser la moyenne et la variabilité d'un procédé.
- **Carte X barre S** : choisissez cette option si vous voulez calculer une carte X barre combinée avec une carte S pour analyser la moyenne et la variabilité d'un procédé.
- **Carte X barre S²** : choisissez cette option si vous voulez calculer une carte X barre combinée avec une carte S² pour analyser la moyenne et la variabilité d'un procédé.

Format des données : choisissez le format des données.

- **Colonnes/Lignes** : activez cette option pour que XLSTAT considère chaque colonne (en mode colonne) ou ligne (en mode ligne) comme une mesure séparée qui appartient au même sous-groupe.
- **Une colonne/ligne** : activez cette option si les mesures des différents sous-groupes se suivent dans la même colonne (mode colonne) ou la même ligne (mode ligne). Dans ce cas, pour affecter les mesures aux différents sous-groupes, indiquez le nombre d'observations par sous-groupe dans le cas où il est constant, et si ce nombre varie ou si les observations ne sont pas triées, sélectionnez une colonne ou une ligne indiquant à quel sous-groupe appartient chaque observation.

Données : si le format « Une colonne/ligne » a été choisi, veuillez sélectionner une unique colonne (ou ligne) qui contient toutes les données. Si le format choisi est « Colonnes/lignes », veuillez sélectionner la plage de données comprenant une colonne ou ligne par sous-groupe.

Groupes : si le format « Une colonne/ligne » a été choisi, activez cette option pour ensuite sélectionner une colonne ou ligne comprenant l'identifiant des groupes.

Effectif des sous-groupes : si le format « Une colonne/ligne » a été choisi et si la taille des groupes est constant, alors vous pouvez désactiver l'option « Groupes » et entrer ici la taille commune des groupes.

Phase : activez cette option pour sélectionner ensuite une colonne/ligne indiquant l'identifiant de la phase. A chaque phase, XLSTAT va recalculer la ligne centrale et les limites de contrôle et créer une nouvelle carte.

- **Spécifications différentes** : activez cette option si vous souhaitez rentrer des spécifications propres à chaque phase pour les paramètres de capabilités du procédé. Dans ce cas, dans l'onglet Options, vous devez rentrer une valeur USL, une valeur LSL et une valeur cible pour chaque phase.

Libellés des observations : activez cette option si vous voulez utiliser des libellés d'observations disponibles sur une feuille Excel. Si l'option « Libellés des colonnes » est activée, la première cellule de la sélection doit comprendre un en-tête. Si vous n'activez pas cette option, des libellés seront automatiquement créés (Obs1, Obs2, …).

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des colonnes/lignes : activez cette option si la première ligne (mode colonnes) ou colonne (mode lignes) des données sélectionnées contient des libellés.

Onglet **Options**:

Limite supérieure de contrôle :

- **Bornée** : activez cette option pour entrer la valeur maximale acceptable pour la limite supérieure de contrôle pour le procédé. Cette valeur sera utilisée si la limite supérieure de contrôle calculée dépasse la valeur entrée.

- **Valeur** : entrez la valeur de la limite supérieure de contrôle à utiliser en remplacement de la valeur calculée.

Limite inférieure de contrôle :

- **Bornée** : activez cette option pour entrer la valeur minimale acceptable pour la limite inférieure de contrôle pour le procédé. Cette valeur sera utilisée si la limite inférieure de contrôle calculée dépasse la valeur entrée.

- **Valeur** : entrez la valeur de la limite inférieure de contrôle à utiliser en remplacement de la valeur calculée.

Calculer les capabilités des procédés : activez cette option pour calculer les capabilités des procédés à partir des données (voir la section description pour plus de détails).

- **USL** : veuillez entrer ici la valeur de la limite supérieure de spécification (USL) du procédé.
- **LSL** : veuillez entrer ici la valeur de la limite inférieure de spécification (LSL) du procédé.
- **Cible** : activez cette option pour ajouter la valeur cible du procédé.
- **Intervalle de confiance (%)** : entrez l'étendue en pourcentage de l'intervalle de confiance à utiliser pour le calcul de l'intervalle de confiance autour des capabilités du procédé. Valeur par défaut : 95.

Transformation Box-Cox : activez cette option pour faire une transformation de Box-Cox. Vous pouvez soit imposer une valeur de **Lambda**, soit décider que XLSTAT doit l'**optimiser** (voir la description pour plus de détails).

k Sigma : activez cette option pour entrer la distance entre les limites de contrôle inférieure et supérieure et la ligne centrale de la carte de contrôle, en terme de nombre d'écarts-types. La distance sera fixée à k fois l'écart-type estimé. Des facteurs de correction fournis par Burr (1969) sont appliqués.

alpha : activez cette option pour définir l'intervalle de confiance autour de la ligne centrale de la carte de contrôle. $100 - \alpha\%$ de la distribution se trouve dans les limites de contrôle. Des facteurs de correction fournis par Burr (1969) sont appliqués.

Moyenne : activez cette option pour entrer la valeur de la ligne centrale de la carte de contrôle. Cette valeur est en général calculée à partir d'un historique.

Sigma : activez cette option pour entrer la valeur de l'écart-type du procédé. Cette valeur est en général calculée à partir d'un historique. Si cette option est activée, vous ne pouvez pas choisir une méthode d'estimation de sigma dans l'onglet "Estimation".

Onglet **Estimation** :

Méthode pour Sigma : choisissez la méthode utilisée pour l'estimation de l'écart-type de la carte de contrôle (voir la description pour plus de détails) :

- Ecart-type global
- R-barre
- S-barre

Onglet **Sorties** :

Afficher les zones : activez cette option pour afficher en plus des limites de contrôle, les limites des zones A et B.

Tests de normalité : activez cette option pour tester la normalité des données (voir l'outil Tests de normalité pour plus de détails).

Niveau de signification (%) : entrez le niveau de signification à utiliser pour les tests (valeur par défaut : 5%).

Test des causes spéciales : activez cette option pour analyser les points de la carte de contrôle en utilisant les règles pour les causes spéciales. Vous pouvez utiliser les règles suivantes :

- 1 point au-delà de 3s de la ligne centrale
- 9 points consécutifs du même côté de la ligne centrale
- 6 points consécutifs tous montant ou tous descendant
- 14 points consécutifs alternant au-dessus et en-dessous
- 2 sur 3 points > 2s de la ligne centrale (du même côté)
- 4 sur 5 points > 1s de la ligne centrale (du même côté)
- 15 points consécutifs plus proche que 1s de la ligne centrale (des deux côtés)
- 8 points > 1s de la ligne centrale (des deux côtés)
- **Toutes** : cliquer sur ce bouton pour sélectionner toutes les options.
- **None** : cliquer sur ce bouton pour désélectionner toutes les options.

Appliquer les règles de Westgard : activez cette option pour analyser les différents points de la carte de contrôle en utilisant les règles de Westgard. Vous pouvez choisir parmi les règles suivantes :

- Règle 1 2s
- Règle 1 3
- Règle 2 2s
- Règle 4s
- Règle 4 1s
- Règle 10 X
- **Toutes** : cliquer sur ce bouton pour sélectionner toutes les options.

- **None** : cliquer sur ce bouton pour désélectionner toutes les options.

Onglet **Graphiques** :

Afficher les graphiques : activez cette option pour visualiser les cartes de contrôle sous forme de graphiques.

Graphiques Q-Q (loi normale) : activez cette option pour afficher des graphiques Q-Q basés sur la loi normale.

Histogrammes : activez cette option pour afficher les histogrammes des échantillons. Pour la distribution théorique, la fonction de densité est affichée.

Run chart : activez cette option pour afficher un graphique figurant les observations de chacun des sous groupes.

Résultats

Estimation :

Moyenne estimée : dans ce tableau sont affichées les moyennes estimées pour les différentes phases.

Ecart-type estimé : dans ce tableau sont affichés les écarts-types estimés pour les différentes phases.

Transformation Box-Cox :

Lambda : ce tableau n'est affiché que si l'option d'optimisation de Lambda a été choisie.

Série avant et après transformation : dans ce tableau sont affichées la série avant transformation et la série après transformation. Si Lambda a été optimisé, la série après optimisation correspond aux résidus du modèle. Si Lambda est fixé, la série après transformation correspond à l'application directe de la transformation de Box-Cox.

Capabilités du procédé :

Capabilités du processus : ces tableaux sont affichés si l'option "Capabilités des procédés" est activée. Il y a un tableau par phase. Un tableau comprend les indicateurs de capacité du processus et si possible les intervalles de confiance correspondant : Cp, Cpl, Cpu, Cpk, Pp, Ppl, Ppu, Ppk, Cpm, Cpm (Boyle), Cp 5.5, Cpk 5.5, Cpmk, et Cs (Wright).

Pour les Cp, Cpl, et Cpu, une information concernant la performance du procédé est fournie, et pour le Cp une information sur la situation est donnée pour faciliter l'interprétation.

Aux valeurs de Cp sont associées les états suivants selon Ekvall et Juran (1974):

- "pas adéquat" si $Cp < 1$
- "adéquat" si $1 \leq Cp \leq 1.33$
- "plus qu'adéquat" si $Cp > 1.33$

D'après Montgomery (2001), le Cp doit avoir les valeurs minimales suivantes pour que la performance du procédé soit comme attendue :

- 1.33 pour les procédés existants
- 1.50 pour de nouveaux procédés ou des procédés existants si la variable est critique
- 1.67 pour de nouveaux procédés si la variable est critique

D'après Montgomery (2001), le Cpu et le Cpl doivent avoir les valeurs minimales suivantes pour que la performance du procédé soit comme attendue :

- 1.25 pour les procédés existants
- 1.45 pour de nouveaux procédés ou des procédés existants si la variable est critique
- 1.60 pour de nouveaux procédés si la variable est critique

Capacités : ce graphique présente l'information concernant les spécifications et les limites de contrôle. La ligne qui joint les limites inférieures et supérieures correspond aux limites inférieures et supérieures, tandis que le barre verticale correspond à la ligne centrale. Les limites de contrôle correspondant aux différentes phases sont affichées séparément.

Graphiques :

Les éléments suivants sont affichés pour chaque graphique sélectionné. Les résultats ci-dessous sont affichés séparément pour chaque carte demandée. Une carte peut être choisie seule ou en combinaison avec la carte X barre.

Carte X barre/ R/ S/ S² : dans ce tableau sont contenues les informations concernant la ligne centrale et les limites de contrôle du graphique en question. Une colonne est affichée pour chaque phase.

Détails pour les observations : dans ce tableau sont affichées des informations détaillées pour chaque sous-groupe. Pour chaque sous-groupe, sont affichés, la phase correspondante, l'effectif, la moyenne, les valeurs minimales et maximales, la ligne centrale et les limites de contrôle inférieure et supérieure. Si l'information concernant les zones A, B et C est activée, alors les limites de contrôle inférieure et supérieure des zones A et B sont aussi affichées.

Détails pour les règles ² : si l'option pour l'application des règles est active, un tableau détaillé concernant les règles est affiché. Pour chaque sous-groupe, il y a une ligne pour chaque règle à appliquer. "Oui" indique que la règle en question a été appliquée pour le sous-groupe correspondant, et "Non" indique que la règle ne s'applique pas.

Carte X barre/ R/ S/ S² : si l'option d'affichage des graphiques est active, alors un graphique construit à partir des tableaux mentionnés ci-dessus est affiché. Chaque sous-groupe est affiché. La ligne centrale ainsi que les limites de contrôle inférieures et supérieures sont aussi affichées. Si les options correspondantes ont été activées, les limites de contrôle des zones A et B sont incluses, et des libellés sont affichés pour les sous-groupes pour lesquels les règles sont appliquées. Une légende comprenant les règles appliquées est affichée sous le graphique.

Tests de normalité :

Pour chaque test demandé sont affichées les statistiques relatives au test, dont notamment la p-value qui est ensuite utilisée pour l'interprétation du test par comparaison avec le seuil de signification choisi.

S'ils ont été demandés, les graphiques P-P et Q-Q sont ensuite affichés.

Histogrammes : les histogrammes sont affichés. Si vous le souhaitez, vous pouvez modifier la couleur des lignes, les échelles, et les titres comme avec n'importe quel graphique Excel.

Run chart : le graphique des derniers points est affiché.

Exemple

Un tutoriel expliquant comment utiliser les cartes de contrôle pour les sous-groupes est disponible sur le Centre d'aide XLSTAT :

http://www.xlstat.com/demoSPS_FR.htm

Bibliographie

Boyles R.A. (1991). The Taguchi capability index. *Journal of Quality Technology*, **23**, 17–26.

Burr I. W. (1967). The effect of non-normality on constants for X and R charts. *Industrial Quality control*, **23(11)**, 563-569.

Burr I. W. (1969). Control charts for measurements with varying sample sizes. *Journal of Quality Technology*, **1(3)**, 163-167.

Deming W. E. (1993). The New Economics for Industry, Government, and Education. Cambridge, MA: Center for Advanced Engineering Study, Massachusetts Institute of Technology.

Ekvall D. N. (1974). Manufacturing Planning. In Quality Control Handbook,. 3rd Ed. (J. M. Juran, et al. eds.) pp. 9-22-39, McGraw-Hill Book Co., New York.

Montgomery D.C. (2001). Introduction to Statistical Quality Control, 4th edition, John Wiley & Sons.

Nelson L.S. (1984). The Shewhart Control Chart - Tests for Special Causes. *Journal of Quality Technology*, **16**, 237-239.

Pyzdek Th. (2003). The Six Sigma Handbook Revised and Expanded, McGraw Hill, New York.

Ryan Th. P. (2000). Statistical Methods for Quality Improvement, Second Edition, Wiley Series in probability and statistics, John Wiley & Sons, New York.

Shewhart W. A. (1931). Economic Control of Quality of Manufactured Product, Van Nostrand, New York.

Wright, P.A. (1995). A process capability index sensitive to skewness. *Journal of Statistical Computation and Simulation*, **52**, 195–203.

Cartes pour valeurs individuelles

Utilisez cet outil pour maîtriser la qualité de vos procédés dans le cas où vous disposez d'une unique mesure pour chaque pas de temps. Les mesures doivent être des données numériques. Cette méthode est utile pour résumer l'évolution de la moyenne et de la variabilité de la qualité d'une production.

Vous trouverez intégrés à cet outil, les transformations Box-Cox et le calcul de capacité du processus, ainsi que la possibilité d'appliquer des règles spéciales ou des règles de Westgard (un ensemble de règles alternatives pour identifier des causes spéciales) pour compléter votre analyse.

Dans cette section :

Description

Les cartes de contrôle ont été d'abord mentionnées par Walter Shewhart dans un document écrit alors qu'il travaillait aux Bell Labs en 1924. Il a ensuite décrit ses méthodes plus en détails dans un livre (1931).

Pendant plusieurs années, il n'y eu pas d'avancées majeures dans ce domaine, jusqu'à ce que Deming mette au point les cartes de contrôle CUSUM, UWMA et EWMA en 1936.

Les cartes de contrôle étaient à l'origine utilisées pour le contrôle de la qualité des biens de production. Pour cette raison, le vocabulaire utilisé pour ces méthodes statistiques provient souvent de ce domaine d'application. Aujourd'hui, ces approches sont appliquées dans de nombreux autres domaines, comme par exemple les services, les ressources humaines ou les ventes. Dans les lignes qui suivent nous utilisons le vocabulaire du domaine de la production.

Cartes pour valeurs individuelles

L'outil de création de cartes de contrôle pour valeurs individuelles permet de créer les graphiques suivants, seuls ou combinés :

- X (individuelle) : une carte X individuelle est utile pour suivre la moyenne mobile d'un procédé de production. Des changements de la moyenne sont aisément repérables sur les cartes.
- EM (étendue mobile) : une carte EM (carte pour l'étendue mobile) est utile pour analyser la variabilité de la production. Des différences importantes de la qualité de la production

dues à des lignes de production différentes sont aisément repérables.

Remarque 1 : si vous souhaitez pouvoir détecter des décalages plus faibles de la moyenne, vous pouvez utiliser les cartes CUSUM qui sont d'ailleurs souvent préférées aux cartes pour valeurs individuelles.

Remarque 2 : si vous ne disposez de plusieurs mesures pour chaque pas de temps, vous devez utiliser les cartes de contrôle pour sous-groupes.

Remarque 3 : si vous mesures sont de nature qualitative (par exemple, oui/non, conforme/non conforme), vous devez utiliser les cartes de contrôle par attributs.

XLSTAT vous propose les options suivantes pour l'estimation de l'écart-type (sigma) d'un échantillon de n mesures :

- Etendue mobile moyenne : sigma est estimé sur la base de l'étendue mobile moyenne avec une fenêtre de m mesures :

$$\hat{s} = \overline{m}/d_2$$

où d_2 est une constante définie par Burr (1969).

- Etendue mobile médiane : sigma est estimé sur la base de l'étendue mobile médiane avec une fenêtre de m mesures :

$$\hat{s} = \overline{median}/d_4$$

où d_4 est une constante définie par Burr (1969).

- S-barre : sigma est calculé à partir des n mesures, selon la formule suivante :

$$\hat{s} = s/c_4,$$

où s est l'écart-type observé sur les n mesures, et où c_4 est une constante définie par Burr (1969).

Capabilité du procédé

La capabilité du procédé décrit un procédé (ou processus) et permet de savoir s'il est sous contrôle et si les données correspondant aux variables mesurées sont à l'intérieur des limites de spécification du procédé. Dans un tel cas, on dit que le procédé est « capable ».

Au cours de l'interprétation des différents indicateurs de capabilité des procédés, veuillez prendre garde au fait que certains indicateurs nécessitent de faire l'hypothèse de normalité ou, tout au moins, de la symétrie de la distribution des variables mesurées. En utilisant les tests de normalité vous pourrez vérifier la validité de ces hypothèses (voir l'outil Tests de Normalité).

Si l'hypothèse de normalité ne peut être retenue, vous avez les possibilités suivantes pour obtenir des capacités des processus :

- Utiliser une transformation Box-Cox pour améliorer la normalité des échantillons, et vérifier ensuite à nouveau la normalité avec un test.

- Utiliser l'indicateur de capacité de processus Cp 5.15.

Soit m, s, LSL, USL respectivement les estimateurs de la moyenne, l'écart-type, la limite de spécification inférieure, la limite de spécification supérieure du procédé, et T la cible choisie. Soit μ et σ les moyenne et écart-type théorique du procédé. XLSTAT permet de calculer les indices de performance suivants pour évaluer la capabilité du procédé :

- Cp : L'indice de capabilité court terme du procédé est estimé par :

$$\hat{C}p = \frac{USL - LSL}{6s}$$

- Cpl : L'indice de capabilité court terme inférieure du procédé est estimé par :

$$\hat{C}pl = \frac{m - LSL}{3s}$$

- Cpu : L'indice de capabilité court terme supérieure du procédé est estimé par :

$$\hat{C}pu = \frac{USL - m}{3s}$$

- Cpk : Cet indice de capabilité court terme, qui contrairement au Cp nécessite la connaissance de la moyenne, est estimé par :

$$\hat{C}pk = min(Cpl, Cpu)$$

- Pp : La capabilité long terme du procédé est définie par:

$$Pp = \frac{USL - LSL}{6\sigma}$$

- Ppl : La capabilité long terme inférieure du procédé est définie par:

$$Ppl = \frac{\mu - LSL}{3\sigma}$$

- Ppu : La capabilité long terme supérieure du procédé est définie par:

$$Ppu = \frac{USL - \mu}{3\sigma}$$

- Ppk : La capabilité long terme Ppk, qui contrairement au Pp nécessite la connaissance de la moyenne, est défini par :

$$Ppk = min(Ppl, Ppu)$$

- Cpm : L'indice de capabilité court terme de Taguchi peut être calculé si une valeur cible (T) a été spécifiée. Son estimateur est défini par :

$$\hat{C}pm = \frac{USL - LSL}{6\sqrt{s^2 + (m - T)^2}}$$

- $CpmBoyles$: L'indice de capabilité court terme de Taguchi amélioré par Boyles (1991) qui nécessite également la spécification d'une valeur cible (T), a pour estimateur :

$$\hat{C}pm\ Boyles = \frac{USL - LSL}{6\sqrt{(n-1)s^2/n + (m - T)^2}}$$

- $Cp5.15$: Cet indice de capabilité court terme est défini par :

$$\hat{C}p\ 5.15 = \frac{USL - LSL}{5.15s}$$

- $Cpk 5.15: $ Cet indice de capabilité court terme est défini par :

$$\hat{C}pk\ 5.15 = \frac{(USL - LSL)/2 - |m - (USL + LSL)/2|}{2.57s}$$

- $Cpmk$: Cet indice de capabilité court terme a été proposé par Pearn. Il peut être calculé si la valeur cible a été spécifiée :

$$\hat{C}pmk = \frac{s \cdot Cpk}{\sqrt{s^2 + (m - T)^2}}$$

- $CsWright$: La capabilité du procédé telle que proposée par Wright (1995). Elle peut être calculée si la valeur cible a été spécifiée :

$$\hat{C}s = \frac{(USL - LSL)/2 - |m - T|}{3\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - T)^2 + \left|\frac{n^2 m_3}{(n-1)(n-2)} \times \left(\frac{n}{n-1} \times \frac{m_2}{c_4^2}\right)^{-1/2}\right|}}$$

avec

$$m_r = \sum_{i=1}^{n}(x_i - T)^r$$

$$c_4 = \left[\frac{2}{n-1}\right]^{1/2} \Gamma(\frac{n}{2})/\Gamma(\frac{n-1}{2})^{-1}$$

- $Zbelow$: Le nombre d'écarts-types entre la moyenne et la limite de spécification inférieure. Il est défini par :

$$Zbelow = (m - LSL)/s$$

- $Zabove$: Le nombre d'écarts-types entre la moyenne et la limite de spécification supérieure. Il est défini par :

$$Zabove = (USL - m)/s$$

- $Ztotal$: Le nombre d'écarts-types entre les limites de spécification inférieure et supérieure. Il est défini par:

$$Ztotal = (USL - LSL)/s$$

- $p(notconform)below$: La probabilité d'avoir un produit défectueux sous la limite de spécification inférieure :

$$p(\text{not conform})below = \Phi(Zbelow)$$

- $p(notconform)above$: La probabilité d'avoir un produit défectueux au-delà de la limite de spécification supérieure :

$$p(\text{not conform})above = \Phi(Zabove)$$

- $p(notconform)total$: La probabilité d'avoir un produit défectueux sous la limite de spécification inférieure ou au-delà de la limite de spécification supérieure :

$$p(\text{not conform})total = p(\text{not conform})below + p(\text{not conform})above$$

- $PPMbelow$: Le nombre de produits défectueux sous la limite de spécification inférieure pour une production d'un million de produits :

$$PPMbelow = p(notconform)below \times 10^6$$

- $PPMabove$: Le nombre de produits défectueux au-delà de la limite de spécification supérieure pour une production d'un million de produits :

$$PPMabove = p(notconform)above \times 10^6$$

- $PPMtotal$: Le nombre de produits défectueux hors des limites de spécification pour une production d'un million de produits :

$$PPMtotal = PPMbelow + PPMabove$$

Transformation Box-Cox

Transformation Box-Cox : activez cette option pour faire une transformation de Box-Cox. Vous pouvez soit imposer une valeur de λ, soit décider que XLSTAT doit l'optimiser. Cette transformation permet d'augmenter la normalité des données ; l'équation de Box-Cox est définie par :

$$Y_t = \begin{cases} \frac{X_t^\lambda - 1}{\lambda}, & X_t \geq 0, \lambda > 0 \\ \ln(X_t), & X_t > 0, \lambda = 0 \end{cases}$$

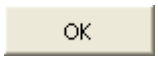
Si l'option d'optimisation est choisie, XLSTAT maximise la vraisemblance de l'échantillon, étant supposé qu'après transformation l'échantillon suit une loi normale.

Règles pour l'interprétation des cartes

XLSTAT vous donne la possibilité d'appliquer des règles pour les « causes spéciales » ainsi que les règles de Westgard. Deux ensembles de règles sont proposées pour l'interprétation des graphiques. Vous pouvez activer ou désactiver les règles dans chacun d'eux.

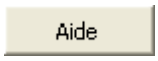
Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.



: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.


: cliquez sur ce bouton pour changer la façon dont vous chargez les données dans XLSTAT. Si le bouton possède une icône de souris, XLSTAT vous permet de faire une sélection à la souris des données. Si le bouton possède une icône de liste, XLSTAT vous permet de sélectionner les données à partir d'une liste. Si le bouton possède une icône orange, des boutons complémentaires avec un point d'interrogation  sont affichés pour vous permettre d'importer les données à partir de fichiers.

Onglet **Général** :

Type de carte : choisissez le type de carte de contrôle que vous voulez utiliser :

- **Carte X individuelle** : choisissez cette option si vous voulez calculer une carte X individuelle pour analyser la moyenne d'un procédé.
- **Carte étendue mobile EM** : choisissez cette option si vous voulez calculer une carte EM pour analyser la variabilité d'un processus.

Données : veuillez sélectionner une unique colonne (ou ligne) qui contient toutes les données.

Phase : activez cette option pour sélectionner ensuite une colonne/ligne indiquant l'identifiant de la phase.

- **Spécifications différentes** : activez cette option si vous souhaitez rentrer des spécifications propres à chaque phase pour les paramètres de capabilités du procédé. Dans ce cas, dans l'onglet Options, vous devez rentrer une valeur USL, une valeur LSL et une valeur cible pour chaque phase.

Libellés des observations : activez cette option si vous voulez utiliser des libellés d'observations disponibles sur une feuille Excel. Si l'option « Libellés des colonnes » est activée, la première cellule de la sélection doit comprendre un en-tête. Si vous n'activez pas cette option, des libellés seront automatiquement créés (Obs1, Obs2, …).

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des colonnes/lignes : activez cette option si la première ligne (mode colonnes) ou colonne (mode lignes) des données sélectionnées contient des libellés.

Onglet **Options** :

Limite supérieure de contrôle :

- **Bornée** : activez cette option pour entrer la valeur maximale acceptable pour la limite supérieure de contrôle pour le procédé. Cette valeur sera utilisée si la limite supérieure de contrôle calculée dépasse la valeur entrée.
- **Valeur** : entrez la valeur de la limite supérieure de contrôle à utiliser en remplacement de la valeur calculée.

Limite inférieure de contrôle :

- **Bornée** : activez cette option pour entrer la valeur minimale acceptable pour la limite inférieure de contrôle pour le procédé. Cette valeur sera utilisée si la limite inférieure de contrôle calculée dépasse la valeur entrée.
- **Valeur** : entrez la valeur de la limite inférieure de contrôle à utiliser en remplacement de la valeur calculée.

Calculer les capabilités des procédés : activez cette option pour calculer les capabilités des procédés à partir des données (voir la section description pour plus de détails).

- **USL** : veuillez entrer ici la valeur de la limite supérieure de spécification (USL) du procédé.

- **LSL** : veuillez entrer ici la valeur de la limite inférieure de spécification (LSL) du procédé.
- **Cible** : activez cette option pour ajouter la valeur cible du procédé.
- **Intervalle de confiance (%)** : entrez l'étendue en pourcentage de l'intervalle de confiance à utiliser pour le calcul de l'intervalle de confiance autour des capabilités du procédé. Valeur par défaut : 95.

Transformation Box-Cox : activez cette option pour faire une transformation de Box-Cox. Vous pouvez soit imposer une valeur de **Lambda**, soit décider que XLSTAT doit l'**optimiser** (voir la section description pour plus de détails).

k Sigma : activez cette option pour entrer la distance entre les limites de contrôle inférieure et supérieure et la ligne centrale de la carte de contrôle, en terme de nombre d'écarts-types. La distance sera fixée à k fois l'écart-type estimé. Des facteurs de correction fournis par Burr (1969) sont appliqués.

alpha : activez cette option pour définir l'intervalle de confiance autour de la ligne centrale de la carte de contrôle. $100 - \alpha\%$ de la distribution se trouve dans les limites de contrôle. Des facteurs de correction fournis par Burr (1969) sont appliqués.

Moyenne : activez cette option pour entrer la valeur de la ligne centrale de la carte de contrôle. Cette valeur est en général calculée à partir d'un historique.

Sigma : activez cette option pour entrer la valeur de l'écart-type du procédé. Cette valeur est en général calculée à partir d'un historique. Si cette option est activée, vous ne pouvez pas choisir une méthode d'estimation de sigma dans l'onglet "Estimation".

Onglet **Estimation** :

Méthode pour Sigma : choisissez la méthode utilisée pour l'estimation de l'écart-type de la carte de contrôle (voir la section description pour plus de détails) :

- Etendue mobile moyenne
- Etendue mobile médiane
- Longueur des EM : changez cette valeur pour modifier le nombre de valeurs prises en compte pour le calcul des étendues mobiles.
- S-barre

Onglet **Sorties** :

Afficher les zones : activez cette option pour afficher en plus des limites de contrôle, les limites des zones A et B.

Tests de normalité : activez cette option pour tester la normalité des données (voir l'outil Tests de normalité pour plus de détails).

Niveau de signification (%) : entrez le niveau de signification à utiliser pour les tests (valeur par défaut : 5%).

Test des causes spéciales : activez cette option pour analyser les points de la carte de contrôle en utilisant les règles pour les causes spéciales. Vous pouvez utiliser les règles suivantes :

- 1 point au-delà de 3s de la ligne centrale
- 9 points consécutifs du même côté de la ligne centrale
- 6 points consécutifs tous montant ou tous descendant
- 14 points consécutifs alternant au-dessus et en-dessous
- 2 sur 3 points > 2s de la ligne centrale (du même côté)
- 4 sur 5 points > 1s de la ligne centrale (du même côté)
- 15 points consécutifs plus proche que 1s de la ligne centrale (des deux côtés)
- 8 points > 1s de la ligne centrale (des deux côtés)
- **Toutes** : cliquer sur ce bouton pour sélectionner toutes les options.
- **None** : cliquer sur ce bouton pour désélectionner toutes les options.

Appliquer les règles de Westgard : activez cette option pour analyser les différents points de la carte de contrôle en utilisant les règles de Westgard. Vous pouvez choisir parmi les règles suivantes :

- Règle 1 2s
- Règle 1 3
- Règle 2 2s
- Règle 4s
- Règle 4 1s
- Règle 10 X
- **Toutes** : cliquer sur ce bouton pour sélectionner toutes les options.
- **None** : cliquer sur ce bouton pour désélectionner toutes les options.

Onglet **Graphiques** :

Afficher les graphiques : activez cette option pour visualiser les cartes de contrôle sous forme de graphiques.

Graphiques Q-Q (loi normale) : activez cette option pour afficher des graphiques Q-Q basés sur la loi normale.

Histogrammes : activez cette option pour afficher les histogrammes des échantillons. Pour la distribution théorique, la fonction de densité est affichée.

Run chart : activez cette option pour afficher un graphique figurant les observations de chacun des sous groupes.

Résultats

Estimation :

Moyenne estimée : dans ce tableau sont affichées les moyennes estimées pour les différentes phases.

Ecart-type estimé : dans ce tableau sont affichés les écarts-types estimés pour les différentes phases.

Transformation Box-Cox :

Lambda : ce tableau n'est affiché que si l'option d'optimisation de Lambda a été choisie.

Série avant et après transformation : dans ce tableau sont affichées la série avant transformation et la série après transformation. Si Lambda a été optimisé, la série après optimisation correspond aux résidus du modèle. Si Lambda est fixé, la série après transformation correspond à l'application directe de la transformation de Box-Cox.

Capabilités du procédé :

Capabilités du procédé : ces tableaux sont affichés si l'option "Capabilités des procédés" est activée. Il y a un tableau par phase. Un tableau comprend les indicateurs de capacité du processus et si possible les intervalles de confiance correspondant : Cp, Cpl, Cpu, Cpk, Pp, Ppl, Ppu, Ppk, Cpm, Cpm (Boyle), Cp 5.5, Cpk 5.5, Cpmk, and Cs (Wright).

Pour les Cp, Cpl, et Cpu, une information concernant la performance du procédé est fournie, et pour le Cp une information sur la situation est donnée pour faciliter l'interprétation.

Aux valeurs de Cp sont associées les états suivants selon Ekvall et Juran (1974):

- "pas adéquat" si $Cp < 1$
- "adéquat" si $1 \leq Cp \leq 1.33$
- "plus qu'adéquat" si $Cp > 1.33$

D'après Montgomery (2001), le Cp doit avoir les valeurs minimales suivantes pour que la performance du procédé soit comme attendue :

- 1.33 pour les procédés existants
- 1.50 pour de nouveaux procédés ou des procédés existants si la variable est critique
- 1.67 pour de nouveaux procédés si la variable est critique

D'après Montgomery (2001), le Cpu et le Cpl doivent avoir les valeurs minimales suivantes pour que la performance du procédé soit comme attendue :

- 1.25 pour les procédés existants
- 1.45 pour de nouveaux procédés ou des procédés existants si la variable est critique
- 1.60 pour de nouveaux procédés si la variable est critique

Capabilités : ce graphique présente l'information concernant les spécifications et les limites de contrôle. La ligne qui joint les limites inférieure et supérieure correspond aux limites inférieure et supérieure, tandis que le barre verticale correspond à la ligne centrale. Les limites de contrôle correspondant aux différentes phases sont affichées séparément.

Graphiques :

Les éléments suivants sont affichés pour chaque graphique sélectionné. Les résultats ci-dessous sont affichés séparément pour chaque carte demandée. Une carte peut être choisie seule ou en combinaison avec la carte X individuelle.

Carte X individuelle / EM : dans ce tableau sont contenues les informations concernant la ligne centrale et les limites de contrôle du graphique en question. Une colonne est affichée pour chaque phase.

Détails pour les observations : dans ce tableau sont affichées des informations détaillées pour chaque sous-groupe. Pour chaque sous-groupe, sont affichés, la phase correspondante, l'effectif, la moyenne, les valeurs minimales et maximales, la ligne centrale et les limites de contrôle inférieure et supérieure. Si l'information concernant les zones A, B et C est activée, alors les limites de contrôle inférieure et supérieure des zones A et B sont aussi affichées.

Détails pour les règles ² : si l'option pour l'application des règles est active, un tableau détaillé concernant les règles est affiché. Pour chaque sous-groupe, il y a une ligne pour chaque règle à appliquer. "Oui" indique que la règle en question a été appliquée, et "Non" indique que la règle ne s'applique pas.

Carte X individuelle / EM : si l'option d'affichage des graphiques est active, alors un graphique construit à partir des tableaux mentionnés ci-dessus est affiché. Chaque observation est affichée. La ligne centrale ainsi que les limites de contrôle inférieures et supérieures sont aussi affichées. Si les options correspondantes ont été activées, les limites de contrôle des zones A

et B sont incluses, et des libellés sont affichés pour les sous-groupes pour lesquels les règles sont appliquées. Une légende comprenant les règles appliquées est affichée sous le graphique.

Tests de normalité :

Pour chaque test demandé sont affichées les statistiques relatives au test, dont notamment la p-value qui est ensuite utilisée pour l'interprétation du test par comparaison avec le seuil de signification choisi.

S'ils ont été demandés, les graphiques P-P et Q-Q sont ensuite affichés.

Histogrammes : les histogrammes sont affichés. Si vous le souhaitez, vous pouvez modifier la couleur des lignes, les échelles, et les titres comme avec n'importe quel graphique Excel.

Run chart : le graphique des derniers points est affiché.

Exemple

Un tutoriel expliquant comment utiliser les cartes de contrôle pour les sous-groupes est disponible sur le Centre d'aide XLSTAT :

http://www.xlstat.com/demoSPI_FR.htm

Bibliographie

Burr I. W. (1967). The effect of non-normality on constants for X and R charts. *Industrial Quality control*, **23(11)**, 563-569.

Burr, I. W. (1969). Control charts for measurements with varying sample sizes. *Journal of Quality Technology*, **1(3)**, 163-167.

Deming, W. E. (1993). The New Economics for Industry, Government, and Education. Cambridge, MA: Center for Advanced Engineering Study, Massachusetts Institute of Technology.

Ekvall D. N. (1974). Manufacturing Planning. In Quality Control Hand-. book,. 3rd Ed. (J. M. **Juran**, et al. eds.) pp. 9-22-39, McGraw-Hill Book Co., New York

Montgomery D.C. (2001). Introduction to Statistical Quality Control, 4th edition, John Wiley & Sons.

Nelson L.S. (1984). The Shewhart Control Chart - Tests for Special Causes. *Journal of Quality Technology*, 16, 237-239.

Pyzdek Th. (2003). The Six Sigma Handbook Revised and Expanded, McGraw Hill, New York.

Ryan Th. P. (2000). Statistical Methods for Quality Improvement, Second Edition, Wiley Series in probability and statistics, John Wiley & Sons, New York.

Shewhart W. A. (1931). Economic Control of Quality of Manufactured Product, Van Nostrand, New York.

Cartes de contrôle par attributs

Utilisez cet outil pour maîtriser la qualité de vos procédés dans le cas où vous disposez d'une unique mesure pour chaque pas de temps. Les mesures sont fondées sur un attribut ou le comptage d'un attribut du procédé. Cette méthode est utile pour résumer l'évolution de variables catégorielles décrivant la qualité d'une production.

Vous trouverez intégrés à cet outil, les transformations Box-Cox et le calcul de capacité du processus, ainsi que la possibilité d'appliquer des règles spéciales ou des règles de Westgard (un ensemble de règles alternatives pour identifier des causes spéciales) pour compléter votre analyse.

Dans cette section :

Description

Les cartes de contrôle ont été d'abord mentionnées par Walter Shewhart dans un document écrit alors qu'il travaillait aux Bell Labs en 1924. Il a ensuite décrit ses méthodes plus en détails dans un livre (1931).

Pendant plusieurs années, il n'y eu pas d'avancées majeures dans ce domaine, jusqu'à ce que Deming mette au point les cartes de contrôle CUSUM, UWMA et EWMA en 1936.

Les cartes de contrôle étaient à l'origine utilisées pour le contrôle de la qualité des biens de production. Pour cette raison, le vocabulaire utilisé pour ces méthodes statistiques provient souvent de ce domaine d'application. Aujourd'hui, ces approches sont appliquées dans de nombreux autres domaines, comme par exemple les services, les ressources humaines ou les ventes. Dans les lignes qui suivent nous utilisons le vocabulaire du domaine de la production.

Cartes par attributs

Ces cartes permettent d'analyser des « produits non conformes » ou des « non conformités ». Ils sont utilisés pour contrôler la qualité avant livraison (produits fabriqués) ou la qualité à la réception (produits achetés). Tous les produits ne sont pas nécessairement contrôlés.

Les inspections sont effectuées par unités d'inspection de taille bien définie. La taille peut être 1 s'il s'agit de télévisions lors de leur réception dans un entrepôt (chaque télévision est inspectée). Elle sera en revanche de 24 dans le cas de cagettes de pêches contenant 24 pêches chacune.

Les différentes cartes par attributs sont les suivantes :

- Carte P : elle est utile pour suivre la proportion d'unités non conformes dans un procédé de production.
- Carte NP : elle est utile pour le nombre absolu d'unités non conformes dans un procédé de production.
- Carte C : elle est utile dans le cas d'une production pour laquelle le nombre d'unités inspectées est constant pour chaque unité inspectée. Elle permet de suivre dans le temps le nombre absolu d'unités non conformes pour chaque contrôle.
- Carte U : elle est utile dans le cas d'une production pour laquelle le nombre d'unités inspectées n'est pas constant. Elle permet de suivre dans le temps la proportion d'unités non conformes pour chaque contrôle.

Les cartes P et NP permettent d'analyser la proportion, respectivement le nombre absolu, de produits non conformes dans un procédé de production. Par exemple, on pourrait compter le nombre d'appareils de télévision non conformes, ou le nombre de cagettes qui comportent au moins une pêche abîmée.

Les cartes C et U permettent d'analyser la proportion, respectivement le nombre absolu, d'occurrences de non conformités dans une unité contrôlée. On peut compter le nombre de produits non conformes dans un procédé de production. Par exemple, on pourrait compter le nombre de transistors défectueux dans une unité contrôlée (il peut y avoir plusieurs transistors défectueux dans une télévision), ou le nombre de pêches abîmées par cagette.

Capacité du processus

La capacité du processus décrit un processus (ou procédé) et permet de savoir s'il est sous contrôle et si les données correspondant aux variables mesurées sont à l'intérieur des limites de spécification du procédé. Dans un tel cas, on dit que le procédé est « capable ».

Au cours de l'interprétation des différents indicateurs de capacité des processus, veuillez prendre garde au fait que certains indicateurs nécessitent de faire l'hypothèse de normalité ou, tout au moins, de la symétrie de la distribution des variables mesurées. En utilisant les tests de normalité vous pourrez vérifier la validité de ces hypothèses (voir les Tests de Normalité).

Si l'hypothèse de normalité ne peut être retenue, vous avez les possibilités suivantes pour obtenir des capacités des processus :

- Utiliser une transformation Box-Cox pour améliorer la normalité des échantillons, et vérifier ensuite à nouveau la normalité avec un test.
- Utiliser l'indicateur de capacité de processus Cp 5.5.

Soit m, s, LSL, USL respectivement les estimateurs de la moyenne, l'écart-type, la limite de spécification inférieure, la limite de spécification supérieure du procédé, et T la cible choisie. Soit μ et σ les moyenne et écart-type théorique du procédé. XLSTAT permet de calculer les indices de performance suivants pour évaluer la capabilité du procédé :

- Cp **:** L'indice de capabilité court terme du procédé est estimé par :

$$\hat{C}p = \frac{USL - LSL}{6s}$$

- Cpl : L'indice de capabilité court terme inférieure du procédé est estimé par :

$$\hat{C}pl = \frac{m - LSL}{3s}$$

- Cpu : L'indice de capabilité court terme supérieure du procédé est estimé par :

$$\hat{C}pu = \frac{USL - m}{3s}$$

- Cpk : Cet indice de capabilité court terme, qui contrairement au Cp nécessite la connaissance de la moyenne, est estimé par :

$$\hat{C}pk = min(Cpl, Cpu)$$

- Pp : La capabilité long terme du procédé est définie par :

$$Pp = \frac{USL - LSL}{6\sigma}$$

- Ppl : La capabilité long terme inférieure du procédé est définie par :

$$Ppl = \frac{\mu - LSL}{3\sigma}$$

- Ppu : La capabilité long terme supérieure du procédé est définie par :

$$Ppu = \frac{USL - \mu}{3\sigma}$$

- Ppk : La capabilité long terme Ppk, qui contrairement au Pp nécessite la connaissance de la moyenne, est défini par :

$$Ppk = min(Ppl, Ppu)$$

- Cpm : L'indice de capabilité court terme de Taguchi peut être calculé si une valeur cible (T) a été spécifiée. Son estimateur est défini par :

$$\hat{C}pm = \frac{USL - LSL}{6\sqrt{s^2 + (m - T)^2}}$$

- $CpmBoyles$: L'indice de capabilité court terme de Taguchi amélioré par Boyles (1991) qui nécessite également la spécification d'une valeur cible (T), a pour estimateur :

$$\hat{C}pm\ Boyles = \frac{USL - LSL}{6\sqrt{(n-1)s^2/n + (m - T)^2}}$$

- $Cp5.15$: Cet indice de capabilité court terme est défini par :

$$\hat{C}p\ 5.15 = \frac{USL - LSL}{5.15s}$$

- $Cpk5.15$: Cet indice de capabilité court terme est défini par :

$$\hat{C}pk\ 5.15 = \frac{(USL - LSL)/2 - |m - (USL + LSL)/2|}{2.57s}$$

- $Cpmk$: Cet indice de capabilité court terme a été proposé par Pearn. Elle peut être calculée si la valeur cible a été spécifiée :

$$\hat{C}pmk = \frac{s \cdot Cpk}{\sqrt{s^2 + (m - T)^2}}$$

- $CsWright$: La capabilité du procédé telle que proposée par Wright (1995). Il peut être calculé si la valeur cible a été spécifiée :

$$\hat{C}s = \frac{(USL - LSL)/2 - |m - T|}{3\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - T)^2 + \left|\frac{n^2 m_3}{(n-1)(n-2)} \times \left(\frac{n}{n-1} \times \frac{m_2}{c_4^2}\right)^{-1/2}\right|}}$$

avec

$$m_r = \sum_{i=1}^{n}(x_i - T)^r$$

$$c_4 = \left[\frac{2}{n-1}\right]^{1/2} \Gamma(\frac{n}{2})/\Gamma(\frac{n-1}{2})^{-1}$$

- $Zbelow$: Le nombre d'écarts-types entre la moyenne et la limite de spécification inférieure. Il est défini par :

$$Zbelow = (m - LSL)/s$$

- $Zabove$: Le nombre d'écarts-types entre la moyenne et la limite de spécification supérieure. Il est défini par :

$$Zabove = (USL - m)/s$$

- $Ztotal$: Le nombre d'écarts-types entre les limites de spécification inférieure et supérieure. Il est défini par:

$$Ztotal = (USL - LSL)/s$$

- $p(notconform)below$: La probabilité d'avoir un produit défectueux sous la limite de spécification inférieure :

$$p(\text{not conform})below = \Phi(Zbelow)$$

- $p(notconform)above$ **:** La probabilité d'avoir un produit défectueux au-delà de la limite de spécification supérieure :

$$p(\text{not conform})above = \Phi(Zabove)$$

- $p(notconform)total$ **:** La probabilité d'avoir un produit défectueux sous la limite de spécification inférieure ou au-delà de la limite de spécification supérieure :

$$p(\text{not conform})total = p(\text{not conform})below + p(\text{not conform})above$$

- $PPMbelow$ **:** Le nombre de produits défectueux sous la limite de spécification inférieure pour une production d'un million de produits :

$$PPMbelow = p(notconform)below \times 10^6$$

- $PPMabove$ **:** Le nombre de produits défectueux au-delà de la limite de spécification supérieure pour une production d'un million de produits :

$$PPMabove = p(notconform)above \times 10^6$$

- $PPMtotal$ **:** Le nombre de produits défectueux hors des limites de spécification pour une production d'un million de produits :

$$PPMtotal = PPMbelow + PPMabove$$

Transformation Box-Cox

Cette transformation permet d'augmenter la normalité des données. Vous pouvez soit imposer une valeur de Lambda, soit décider que XLSTAT l'optimise. L'équation de Box-Cox est définie par :

$$Y_t = \begin{cases} \dfrac{X_t^\lambda - 1}{\lambda}, & (X_t > 0, \lambda \neq 0) \text{ ou } (X_t \geq 0, \lambda > 0), \\ ln(X_t), & X_t > 0, \lambda = 0. \end{cases}$$

Si l'option d'optimisation est choisie, XLSTAT maximise la vraisemblance de l'échantillon, étant supposé qu'après transformation l'échantillon suit une loi normale.

Règles pour l'interprétation des cartes

XLSTAT vous donne la possibilité d'appliquer des règles pour les « causes spéciales » ainsi que les règles de Westgard. Deux ensembles de règles sont proposées pour l'interprétation des graphiques. Vous pouvez activer ou désactiver les règles dans chacun d'eux.

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.

OK : cliquez sur ce bouton pour lancer les calculs.

Annuler : cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

Aide : cliquez sur ce bouton pour afficher l'aide.

 : cliquez sur ce bouton pour rétablir les options par défaut.

 : cliquez sur ce bouton pour effacer les sélections de données.

 : cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.

 : cliquez sur ce bouton pour changer la façon dont vous chargez les données dans XLSTAT. Si le bouton possède une icône de souris, XLSTAT vous permet de faire une sélection à la souris des données. Si le bouton possède une icône de liste, XLSTAT vous permet de sélectionner les données à partir d'une liste. Si le bouton possède une icône orange, des boutons complémentaires avec un point d'interrogation  sont affichés pour vous permettre d'importer les données à partir de fichiers.

Onglet **Général** :

Type de carte : choisissez le type de carte de contrôle que vous voulez utiliser (voir la section description pour plus de détails) :

- **Carte P**
- **Carte NP**
- **Carte C**
- **Carte U**

Données : sélectionnez la plage de données comprenant l'unique colonne ou ligne de données.

Groupes : activez cette option pour ensuite sélectionner une colonne ou ligne comprenant l'identifiant des groupes.

Effectif des sous-groupes : si la taille des groupes est constant, alors vous pouvez désactiver l'option « Groupes » et entrer ici la taille commune des groupes.

Phase : activez cette option pour sélectionner ensuite une colonne/ligne indiquant l'identifiant de la phase.

- **Spécifications différentes** : activez cette option si vous souhaitez rentrer des spécifications propres à chaque phase pour les paramètres de capabilités du procédé.

Dans ce cas, dans l'onglet Options, vous devez rentrer une valeur USL, une valeur LSL et une valeur cible pour chaque phase.

Libellés des observations : activez cette option si vous voulez utiliser des libellés d'observations disponibles sur une feuille Excel. Si l'option « Libellés des colonnes » est activée, la première cellule de la sélection doit comprendre un en-tête. Si vous n'activez pas cette option, des libellés seront automatiquement créés (Obs1, Obs2, …).

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des colonnes/lignes : activez cette option si la première ligne (mode colonnes) ou colonne (mode lignes) des données sélectionnées contient des libellés.

Onglet **Options** :

Limite supérieure de contrôle (UCL) :

- **Bornée** : activez cette option pour entrer la valeur maximale acceptable pour la limite supérieure de contrôle pour le procédé. Cette valeur sera utilisée si la limite supérieure de contrôle calculée dépasse la valeur entrée.
- **Valeur** : entrez la valeur de la limite supérieure de contrôle à utiliser en remplacement de la valeur calculée.

Limite inférieure de contrôle (LCL) :

- **Bornée** : activez cette option pour entrer la valeur minimale acceptable pour la limite inférieure de contrôle pour le procédé. Cette valeur sera utilisée si la limite inférieure de contrôle calculée dépasse la valeur entrée.
- **Valeur** : entrez la valeur de la limite inférieure de contrôle à utiliser en remplacement de la valeur calculée.

Calculer les capacités du processus : activez cette option pour calculer les capacités du processus à partir des données (voir la section description pour plus de détails).

- **USL** : veuillez entrer ici la valeur de la limite supérieure de spécification (USL) du procédé.
- **LSL** : veuillez entrer ici la valeur de la limite inférieure de spécification (LSL) du procédé.
- **Cible** : activez cette option pour ajouter la valeur cible du procédé.

- **Intervalle de confiance (%)** : entrez l'étendue en pourcentage de l'intervalle de confiance à utiliser pour le calcul de l'intervalle de confiance autour des capacités du processus. Valeur par défaut : 95.

Transformation Box-Cox : activez cette option pour faire une transformation de Box-Cox. Vous pouvez soit imposer une valeur de **Lambda**, soit décider que XLSTAT doit l'**optimiser** (voir la section description pour plus de détails).

k Sigma : activez cette option pour entrer la distance entre les limites de contrôle inférieure et supérieure et la ligne centrale de la carte de contrôle, en terme de nombre d'écarts-types. La distance sera fixée à k fois l'écart-type estimé. Des facteurs de correction fournis par Burr (1969) sont appliqués.

alpha : activez cette option pour définir l'intervalle de confiance autour de la ligne centrale de la carte de contrôle. $100 - \alpha\%$ de la distribution se trouve dans les limites de contrôle. Des facteurs de correction fournis par Burr (1969) sont appliqués.

Moyenne : activez cette option pour entrer la valeur de la ligne centrale de la carte de contrôle. Cette valeur est en général calculée à partir d'un historique.

Sigma : activez cette option pour entrer la valeur de l'écart-type du procédé. Cette valeur est en général calculée à partir d'un historique. Si cette option est activée, vous ne pouvez pas choisir une méthode d'estimation de sigma dans l'onglet "Estimation".

P barre / C barre / U barre : activez cette option pour entrer la valeur correspondant à la ligne centrale de la carte de contrôle. Cette valeur doit être basée sur des données historiques.

Onglet **Sorties** :

Afficher les zones : activez cette option pour afficher en plus des limites de contrôle, les limites des zones A et B.

Tests de normalité : activez cette option pour tester la normalité des données (voir l'outil Tests de normalité pour plus de détails).

Niveau de signification (%) : entrez le niveau de signification à utiliser pour les tests (valeur par défaut : 5%).

Test des causes spéciales : activez cette option pour analyser les points de la carte de contrôle en utilisant les règles pour les causes spéciales. Vous pouvez utiliser les règles suivantes :

- 1 point au-delà de 3s de la ligne centrale.
- 9 points consécutifs du même côté de la ligne centrale.

- 6 points consécutifs tous montant ou tous descendant.
- 14 points consécutifs alternant au-dessus et en-dessous.
- 2 sur 3 points > 2s de la ligne centrale (du même côté).
- 4 sur 5 points > 1s de la ligne centrale (du même côté).
- 15 points consécutifs plus proche que 1s de la ligne centrale (des deux côtés).
- 8 points > 1s de la ligne centrale (des deux côtés).
- **Toutes** : cliquez sur ce bouton pour sélectionner toutes les options.
- **None** : cliquez sur ce bouton pour désélectionner toutes les options.

Appliquer les règles de Westgard : activez cette option pour analyser les différents points de la carte de contrôle en utilisant les règles de Westgard. Vous pouvez choisir parmi les règles suivantes :

- Règle 1 2s
- Règle 1 3
- Règle 2 2s
- Règle 4s
- Règle 4 1s
- Règle 10 X
- **Toutes** : cliquez sur ce bouton pour sélectionner toutes les options.
- **None** : cliquez sur ce bouton pour désélectionner toutes les options.

Onglet **Graphiques** :

Afficher les graphiques : activez cette option pour visualiser les cartes de contrôle sous forme de graphiques.

Graphiques Q-Q (loi normale) : activez cette option pour afficher des graphiques Q-Q basés sur la loi normale.

Histogrammes : activez cette option pour afficher les histogrammes des échantillons. Pour la distribution théorique, la fonction de densité est affichée.

Run chart : activez cette option pour afficher un graphique figurant les observations de chacun des sous groupes.

Résultats

Estimation :

Moyenne estimée : dans ce tableau sont affichées les moyennes estimées pour les différentes phases.

Ecart-type estimé : dans ce tableau sont affichés les écarts-types estimés pour les différentes phases.

Transformation Box- Cox :

Lambda : ce tableau n'est affiché que si l'option d'optimisation de Lambda a été choisie.

Série avant et après transformation : dans ce tableau sont affichées la série avant transformation et la série après transformation. Si Lambda a été optimisé, la série après optimisation correspond aux résidus du modèle. Si Lambda est fixé, la série après transformation correspond à l'application directe de la transformation de Box-Cox.

Capacités du processus :

Capacités du processus : ces tableaux sont affichés si l'option "Capacités des processus" est activée. Il y a un tableau par phase. Un tableau comprend les indicateurs de capacité du processus et si possible les intervalles de confiance correspondant : Cp, Cpl, Cpu, Cpk, Pp, Ppl, Ppu, Ppk, Cpm, Cpm (Boyle), Cp 5.5, Cpk 5.5, Cpmk, and Cs (Wright).

Pour les Cp, Cpl, et Cpu, une information concernant la performance du procédé est fournie, et pour le Cp une information sur la situation est donnée pour faciliter l'interprétation.

Aux valeurs de Cp sont associées les états suivants selon Ekvall et Juran (1974):

- "pas adéquat" si Cp < 1,
- "adéquat" si 1 <= Cp <= 1.33,
- "plus qu'adéquat" si Cp > 1.33.

D'après Montgomery (2001), le Cp doit avoir les valeurs minimales suivantes pour que la performance du procédé soit comme attendue :

- 1.33 pour les procédés existants,
- 1.50 pour de nouveaux procédés ou des procédés existants si la variable est critique,
- 1.67 pour de nouveaux procédés si la variable est critique.

D'après Montgomery (2001), le Cpu et le Cpl doivent avoir les valeurs minimales suivantes pour que la performance du procédé soit comme attendue :

- 1.25 pour les procédés existants,
- 1.45 pour de nouveaux procédés ou des procédés existants si la variable est critique,
- 1.60 pour de nouveaux procédés si la variable est critique.

Capacités : ce graphique présente l'information concernant les spécifications et les limites de contrôle. La ligne qui joint les limites inférieures et supérieures correspond aux limites inférieures et supérieures, tandis que le barre verticale correspond à la ligne centrale. Les limites de contrôle correspondant aux différentes phases sont affichées séparément.

Graphiques :

Les éléments suivants sont affichés pour chaque graphique sélectionné. Les résultats ci-dessous sont affichés séparément pour chaque carte demandée. Une carte peut être choisie seule ou en combinaison avec la carte X individuelle.

Carte P / NP / C / U : dans ce tableau sont contenues les informations concernant la ligne centrale et les limites de contrôle du graphique en question. Une colonne est affichée pour chaque phase.

Détails pour les observations : dans ce tableau sont affichées des informations détaillées pour chaque sous-groupe. Pour chaque sous-groupe, sont affichés, la phase correspondante, l'effectif, la moyenne, les valeurs minimales et maximales, la ligne centrale et les limites de contrôle inférieure et supérieure. Si l'information concernant les zones A, B et C est activée, alors les limites de contrôle inférieure et supérieure des zones A et B sont aussi affichées.

Détails pour les règles ² : si l'option pour l'application des règles est active, un tableau détaillé concernant les règles est affiché. Pour chaque sous-groupe, il y a une ligne pour chaque règle à appliquer. "Oui" indique que la règle en question a été appliquée, et "Non" indique que la règle ne s'applique pas.

Carte P / NP / C / U : si l'option d'affichage des graphiques est active, alors un graphique construit à partir des tableaux mentionnés ci-dessus est affiché. La ligne centrale ainsi que les limites de contrôle inférieures et supérieures sont aussi affichées. Si les options correspondantes ont été activées, les limites de contrôle des zones A et B sont incluses, et des libellés sont affichés pour les sous-groupes pour lesquels les règles sont appliquées. Une légende comprenant les règles appliquées est affichée sous le graphique.

Tests de normalité :

Pour chaque test demandé sont affichées les statistiques relatives au test, dont notamment la p-value qui est ensuite utilisée pour l'interprétation du test par comparaison avec le seuil de signification choisi.

S'ils ont été demandés, les graphiques P-P et Q-Q sont ensuite affichés.

Histogrammes : les histogrammes sont affichés. Si vous le souhaitez, vous pouvez modifier la couleur des lignes, les échelles, et les titres comme avec n'importe quel graphique Excel.

Run chart : le graphique des derniers points est affiché.

Exemple

Un tutoriel expliquant comment utiliser les cartes de contrôle par attributs est disponible sur le Centre d'aide XLSTAT sur :

http://www.xlstat.com/demoSPA_FR.htm

Bibliographie

Burr, I. W. (1967). The effect of non-normality on constants for X and R charts. *Industrial Quality control*, 23(11), 563-569.

Burr I. W. (1969). Control charts for measurements with varying sample sizes. *Journal of Quality Technology*, 1(3), 163-167.

Deming W. E. (1993). The New Economics for Industry, Government, and Education. Cambridge, MA: Center for Advanced Engineering Study, Massachusetts Institute of Technology.

Ekvall D. N. (1974). Manufacturing Planning. In Quality Control Handbook,. 3rd Ed. (J. M. Juran, et al. eds.) pp. 9-22-39, McGraw-Hill Book Co., New York.

Montgomery D.C. (2001), Introduction to Statistical Quality Control, 4th edition, John Wiley & Sons.

Nelson L.S. (1984). The Shewhart Control Chart - Tests for Special Causes. *Journal of Quality Technology*, 16, 237-239.

Pyzdek Th. (2003). The Six Sigma Handbook Revised and Expanded, McGraw Hill, New York.

Ryan Th. P. (2000). Statistical Methods for Quality Improvement, Second Edition, Wiley Series in probability and statistics, John Wiley & Sons, New York.

Shewhart W. A. (1931). Economic Control of Quality of Manufactured Product, Van Nostrand, New York.

Cartes de contrôle pondérées par le temps

Utilisez cet outil pour maîtriser la qualité de vos procédés dans le cas où vous avez plusieurs mesures pour chaque pas de temps. Les mesures doivent être des données numériques. Cette méthode est utile pour résumer l'évolution de la moyenne et de la variabilité de la qualité d'une production.

Vous trouverez intégrés à cet outil, les transformations Box-Cox et le calcul de capacité du processus, ainsi que la possibilité d'appliquer des règles spéciales ou des règles de Westgard (un ensemble de règles alternatives pour identifier des causes spéciales) pour compléter votre analyse.

Dans cette section :

Description

Les cartes de contrôle ont été d'abord mentionnées par Walter Shewhart dans un document écrit alors qu'il travaillait aux Bell Labs en 1924. Il a ensuite décrit ses méthodes plus en détails dans un livre (1931).

Pendant plusieurs années, il n'y eu pas d'avancées majeures dans ce domaine, jusqu'à ce que Deming mette au point les cartes de contrôle CUSUM, UWMA et EWMA en 1936.

Les cartes de contrôle étaient à l'origine utilisées pour le contrôle de la qualité des biens de production. Pour cette raison, le vocabulaire utilisé pour ces méthodes statistiques provient souvent de ce domaine d'application. Aujourd'hui, ces approches sont appliquées dans de nombreux autres domaines, comme par exemple les services, les ressources humaines ou les ventes. Dans les lignes qui suivent nous utilisons le vocabulaire du domaine de la production.

Cartes de contrôle pondérées par le temps

Cet outil vous permet de créer les cartes de contrôle suivantes :

- CUSUM ou CUSUM pour valeurs individuelles
- UWMA ou UWMA pour valeurs individuelles
- EWMA ou EWMA pour valeurs individuelles

Une carte CUSUM, UWMA ou EWMA est utile pour suivre la moyenne d'un procédé de fabrication. Les décalages de la moyenne sont aisément repérables sur ces cartes.

Cartes UWMA et EWMA

Ces cartes n'utilisent pas directement les données brutes. Elles sont basées sur des données lissées.

Dans le cas de cartes UWMA, les données sont lissées en utilisant une pondération uniforme sur la fenêtre glissante. La carte est ensuite créée comme une carte Shewhart.

Dans le cas de cartes EWMA, les données sont lissées en utilisant un lissage exponentiel. La carte est ensuite créée comme une carte Shewhart.

Cartes CUSUM

Ces cartes n'utilisent pas directement les données brutes. Elles sont basées sur des données normalisées.

Ces cartes permettent de détecter des décalages de la moyenne avec une granularité définie par l'utilisateur, au travers du paramètre k. k est la moitié du décalage de la moyenne à détecter. Pour détecter un décalable de 1 sigma, k est fixé à 0.5.

On distingue deux type de cartes CUSUM : unilatérale ou bilatérale. Dans le cas d'une carte CUSUM unilatérale les sommes cumulées supérieures et inférieures SH et SL sont calculées récursivement.

$$SH_i = max(0, (z_i - k) + SH_{i-1})$$

$$SL_i = min(0, (z_i + k) + SL_{i-1})$$

Si SH ou SL sont supérieures à un seuil h, alors un décalage est détecté. La valeur de h peut être choisie par l'utilisateur. Les valeurs habituelles sont 4 ou 5.

La valeur initiale de SH et SL au début des calculs ou après la détection d'un décalage est habituellement 0. En utilisant l'option FIR (Fast Initial Response) on peut changer cette valeur de départ. L'utilisateur peut alors entrer la valeur de son choix.

Dans le cas d'une carte CUSUM bilatérale, les données sont normalisées. Les limites de contrôle supérieure et inférieure sont appelées respectivement masque U et masque V. Ces noms sont liés à la forme que prennent les limites de contrôle sur la carte. Pour un point donné, les limites maximales supérieure et inférieure pour la détection d'un décalage, sont calculées en remontant dans le temps et affichées sur la carte avec le format du masque U ou V. Le point pour l'origine du masque est par défaut le dernier point. L'utilisateur peut modifier cela avec l'option « origine ».

XLSTAT vous propose les options suivantes pour l'estimation de l'écart-type (sigma) d'un échantillon, pour k sous-groupes et n_i $(i = 1, \ldots, k)$ mesures par sous-groupe :

- **Ecart-type global** : sigma est calculé sur toutes les mesures disponibles. A partir des k variances intra-sous-groupes, selon la formule suivante :

$$\hat{s} = \sqrt{\frac{\sum_{i=1}^{k}(n_i - 1)\, s_i^2}{\sum_{i=1}^{k}(n_i - 1)}} / c_4 \left(1 + \sum_{i=1}^{k}(n_i - 1)\right)$$

où c_4 est une constante définie par Burr (1969).

- **R-barre** : l'estimateur de sigma est calculé à partir de l'étendue moyenne (ou amplitude moyenne) pour les n sous-groupes. $\hat{S} = \overline{R}/d_2$, où d_2 est une constante définie par Burr (1969).

- **S-barre** : sigma est calculé à partir de la moyenne des k variances intra- sous-groupes, selon la formule suivante : $\hat{s} = \sqrt{\frac{1}{k}\sum_{i=1}^{k} s_i^2}/c_4$ où c_4 est une constante définie par Burr (1969).

Dans le cas de n mesures individuelles :

- **Etendue mobile moyenne** : sigma est estimé sur la base de l'étendue mobile moyenne avec une fenêtre de m mesures : $\hat{s} = \overline{m}/d_2$ où d_2 est une constante définie par Burr (1969).

- **Etendue mobile médiane** : sigma est estimé sur la base de l'étendue mobile médiane avec une fenêtre de m mesures : $\hat{s} = \overline{median}/d_4$, où d_4 est une constante définie par Burr (1969).

- **Ecart-type** : sigma est calculé à partir des n mesures, selon la formule suivante : $\hat{s} = s/c_4$ où s est l'écart-type observé sur les n mesures, et où c_4 est une constante définie par Burr (1969).

Capacité du processus

La capacité du processus décrit un processus (ou procédé) et permet de savoir s'il est sous contrôle et si les données correspondant aux variables mesurées sont à l'intérieur des limites de spécification du procédé. Dans un tel cas, on dit que le procédé est « capable ».

Au cours de l'interprétation des différents indicateurs de capacité des processus, veuillez prendre garde au fait que certains indicateurs nécessitent de faire l'hypothèse de normalité ou, tout au moins, de la symétrie de la distribution des variables mesurées. En utilisant les tests de normalité vous pourrez vérifier la validité de ces hypothèses (voir les Tests de Normalité).

Si l'hypothèse de normalité ne peut être retenue, vous avez les possibilités suivantes pour obtenir des capacités des processus :

- Utiliser une transformation Box-Cox pour améliorer la normalité des échantillons, et vérifier ensuite à nouveau la normalité avec un test.

- Utiliser l'indicateur de capacité de processus Cp 5.5.

Soit m, s, LSL, USL respectivement les estimateurs de la moyenne, l'écart-type, la limite de spécification inférieure, la limite de spécification supérieure du procédé, et T la cible choisie. Soit μ et σ les moyenne et écart-type théorique du procédé. XLSTAT permet de calculer les indices de performance suivants pour évaluer la capabilité du procédé :

- Cp : L'indice de capabilité court terme du procédé est estimé par :

$$\hat{C}p = \frac{USL - LSL}{6s}$$

- Cpl : L'indice de capabilité court terme inférieure du procédé est estimé par :

$$\hat{C}pl = \frac{m - LSL}{3s}$$

- Cpu : L'indice de capabilité court terme supérieure du procédé est estimé par :

$$\hat{C}pu = \frac{USL - m}{3s}$$

- Cpk : Cet indice de capabilité court terme, qui contrairement au Cp nécessite la connaissance de la moyenne, est estimé par :

$$\hat{C}pk = min(Cpl, Cpu)$$

- Pp : La capabilité long terme du procédé est définie par:

$$Pp = \frac{USL - LSL}{6\sigma}$$

- Ppl : La capabilité long terme inférieure du procédé est définie par:

$$Ppl = \frac{\mu - LSL}{3\sigma}$$

- Ppu : La capabilité long terme supérieure du procédé est définie par:

$$Ppu = \frac{USL - \mu}{3\sigma}$$

- Ppk : La capabilité long terme Ppk, qui contrairement au Pp nécessite la connaissance de la moyenne, est défini par :

$$Ppk = min(Ppl, Ppu)$$

- Cpm : L'indice de capabilité court terme de Taguchi peut être calculé si une valeur cible (T) a été spécifiée. Son estimateur est défini par :

$$\hat{C}pm = \frac{USL - LSL}{6\sqrt{s^2 + (m - T)^2}}$$

- $CpmBoyles$: L'indice de capabilité court terme de Taguchi amélioré par Boyles (1991) qui nécessite également la spécification d'une valeur cible (T), a pour estimateur :

$$\hat{C}pm\ Boyles = \frac{USL - LSL}{6\sqrt{(n-1)s^2/n + (m - T)^2}}$$

- $Cp5.15$: Cet indice de capabilité court terme est défini par :

$$\hat{C}p\ 5.15 = \frac{USL - LSL}{5.15s}$$

- $Cpk5.15$: Cet indice de capabilité court terme est défini par :

$$\hat{C}pk\ 5.15 = \frac{(USL - LSL)/2 - |m - (USL + LSL)/2|}{2.57s}$$

- $Cpmk$: Cet indice de capabilité court terme a été proposé par Pearn. Elle peut être calculée si la valeur cible a été spécifiée :

$$\hat{C}pmk = \frac{s \cdot Cpk}{\sqrt{s^2 + (m - T)^2}}$$

- $CsWright$: La capabilité du procédé telle que proposée par Wright (1995). Il peut être calculé si la valeur cible a été spécifiée :

$$\hat{C}s = \frac{(USL - LSL)/2 - |m - T|}{3\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - T)^2 + \left|\frac{n^2 m_3}{(n-1)(n-2)} \times \left(\frac{n}{n-1} \times \frac{m_2}{c_4^2}\right)^{-1/2}\right|}}$$

avec

$$m_r = \sum_{i=1}^{n}(x_i - T)^r$$

$$c_4 = \left[\frac{2}{n-1}\right]^{1/2} \Gamma(\frac{n}{2})/\Gamma(\frac{n-1}{2})^{-1}$$

- $Zbelow$: Le nombre d'écarts-types entre la moyenne et la limite de spécification inférieure. Il est défini par :

$$Zbelow = (m - LSL)/s$$

- $Zabove$: Le nombre d'écarts-types entre la moyenne et la limite de spécification supérieure. Il est défini par :

$$Zabove = (USL - m)/s$$

- $Ztotal$: Le nombre d'écarts-types entre les limites de spécification inférieure et supérieure. Il est défini par:

$$Ztotal = (USL - LSL)/s$$

- $p(notconform)below$: La probabilité d'avoir un produit défectueux sous la limite de spécification inférieure :

$$p(\text{not conform})below = \Phi(Zbelow)$$

- $p(notconform)above$: La probabilité d'avoir un produit défectueux au-delà de la limite de spécification supérieure :

$$p(\text{not conform})above = \Phi(Zabove)$$

- $(notconform)total$: La probabilité d'avoir un produit défectueux sous la limite de spécification inférieure ou au-delà de la limite de spécification supérieure :

$$p(\text{not conform})total = p(\text{not conform})below + p(\text{not conform})above$$

- $PPMbelow$: Le nombre de produits défectueux sous la limite de spécification inférieure pour une production d'un million de produits :

$$PPMbelow = p(notconform)below \times 10^6$$

- $PPMabove$: Le nombre de produits défectueux au-delà de la limite de spécification supérieure pour une production d'un million de produits :

$$PPMabove = p(notconform)above \times 10^6$$

- $PPMtotal$: Le nombre de produits défectueux hors des limites de spécification pour une production d'un million de produits :

$$PPMtotal = PPMbelow + PPMabove$$

Transformation Box-Cox

Transformation Box-Cox : activez cette option pour faire une transformation de Box-Cox. Vous pouvez soit imposer une valeur de λ, soit décider que XLSTAT doit l'optimiser. Cette transformation permet d'augmenter la normalité des données ; l'équation de Box-Cox est définie par :

$$Y_t = \begin{cases} \frac{X_t^\lambda - 1}{\lambda}, & (X_t > 0, \lambda \neq 0) \text{ ou } (X_t \geq 0, \lambda > 0) \\ \ln(X_t), & X_t > 0, \quad \lambda = 0 \end{cases}$$

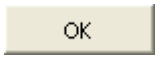
Si l'option d'optimisation est choisie, XLSTAT maximise la vraisemblance de l'échantillon, étant supposé qu'après transformation l'échantillon suit une loi normale.

Règles pour l'interprétation des cartes

XLSTAT vous donne la possibilité d'appliquer des règles pour les « causes spéciales » ainsi que les règles de Westgard. Deux ensembles de règles sont proposées pour l'interprétation des graphiques. Vous pouvez activer ou désactiver les règles dans chacun d'eux.

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.


: cliquez sur ce bouton pour changer la façon dont vous chargez les données dans XLSTAT. Si le bouton possède une icône de souris, XLSTAT vous permet de faire une sélection à la souris des données. Si le bouton possède une icône de liste, XLSTAT vous permet de sélectionner les données à partir d'une liste. Si le bouton possède une icône orange, des boutons complémentaires avec un point d'interrogation  sont affichés pour vous permettre d'importer les données à partir de fichiers.

Onglet **Général** :

Type de carte : choisissez le type de carte de contrôle que vous voulez utiliser (voir la section description pour plus de détails):

- Carte CUSUM
- Carte CUSUM pour valeurs individuelles
- Carte UWMA
- Carte UWMA pour valeurs individuelles
- Carte EWMA

- Carte EWMA pour valeurs individuelles

Format des données : choisissez le format des données.

- **Colonnes/Lignes** : activez cette option pour que XLSTAT considère chaque colonne (en mode colonne) ou ligne (en mode ligne) comme une mesure séparée qui appartient au même sous-groupe.
- **Une colonne/ligne** : activez cette option si les mesures des différents sous-groupes se suivent dans la même colonne (mode colonne) ou la même ligne (mode ligne). Dans ce cas, pour affecter les mesures aux différents sous-groupes, indiquez le nombre d'observations par sous-groupe dans le cas où il est constant, et si ce nombre varie ou si les observations ne sont pas triées, sélectionnez une colonne ou une ligne indiquant à quel sous-groupe appartient chaque observation.

Données : si le format « Une colonne/ligne » a été choisi, veuillez sélectionner une unique colonne (ou ligne) qui contient toutes les données. Si le format choisi est « Colonnes/lignes », veuillez sélectionner la plage de données comprenant une colonne ou ligne par sous-groupe.

Groupes : si le format « Une colonne/ligne » a été choisi, activez cette option pour ensuite sélectionner une colonne ou ligne comprenant l'identifiant des groupes.

Effectif des sous-groupes : si le format « Une colonne/ligne » a été choisi et si la taille des groupes est constant, alors vous pouvez désactiver l'option « Groupes » et entrer ici la taille commune des groupes.

Phase : activez cette option pour sélectionner ensuite une colonne/ligne indiquant l'identifiant de la phase.

- **Spécifications différentes** : activez cette option si vous souhaitez rentrer des spécifications propres à chaque phase pour les paramètres de capabilités du procédé. Dans ce cas, dans l'onglet Options, vous devez rentrer une valeur USL, une valeur LSL et une valeur cible pour chaque phase.

Libellés des observations : activez cette option si vous voulez utiliser des libellés d'observations disponibles sur une feuille Excel. Si l'option « Libellés des colonnes » est activée, la première cellule de la sélection doit comprendre un en-tête. Si vous n'activez pas cette option, des libellés seront automatiquement créés (Obs1, Obs2, …).

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des colonnes/lignes : activez cette option si la première ligne (mode colonnes) ou colonne (mode lignes) des données sélectionnées contient des libellés.

Normaliser : dans le cas d'une carte CUSUM, veuillez activer cette option pour afficher les sommes cumulées et les limites de contrôle normalisées.

Cible : dans le cas d'une carte CUSUM, veuillez activer cette option pour entrer la valeur cible utiliser pour la normalisation des données. La valeur par défaut est la moyenne observée.

Poids : dans le cas d'une carte EWMA, veuillez activer cette option pour entrer le facteur de pondération pour le lissage exponentiel.

Longueur MM : dans le cas d'une carte UWMA, veuillez activer cette option pour entrer la facteur de pondération pour le lissage exponentiel.

Onglet **Options** :

Limite supérieure de contrôle :

- **Bornée** : activez cette option pour entrer la valeur maximale acceptable pour la limite supérieure de contrôle pour le procédé. Cette valeur sera utilisée si la limite supérieure de contrôle calculée dépasse la valeur entrée.
- **Valeur** : entrez la valeur de la limite supérieure de contrôle à utiliser en remplacement de la valeur calculée.

Limite inférieure de contrôle :

- **Bornée** : activez cette option pour entrer la valeur minimale acceptable pour la limite inférieure de contrôle pour le procédé. Cette valeur sera utilisée si la limite inférieure de contrôle calculée dépasse la valeur entrée.
- **Valeur** : entrez la valeur de la limite inférieure de contrôle à utiliser en remplacement de la valeur calculée.

Calculer les capacités du processus : activez cette option pour calculer les capacités du processus à partir des données (voir la section description pour plus de détails).

- **USL** : veuillez entrer ici la valeur de la limite supérieure de spécification (USL) du procédé.
- **LSL** : veuillez entrer ici la valeur de la limite inférieure de spécification (LSL) du procédé.
- **Cible** : activez cette option pour ajouter la valeur cible du procédé.
- **Intervalle de confiance (%)** : entrez l'étendue en pourcentage de l'intervalle de confiance à utiliser pour le calcul de l'intervalle de confiance autour des capacités du processus. Valeur par défaut : 95.

Transformation Box-Cox : activez cette option pour faire une transformation de Box-Cox. Vous pouvez soit imposer une valeur de λ, soit décider que XLSTAT doit l'**optimiser** (voir la section description pour plus de détails).

k Sigma : activez cette option pour entrer la distance entre les limites de contrôle inférieure et supérieure et la ligne centrale de la carte de contrôle, en terme de nombre d'écarts-types. La distance sera fixée à k fois l'écart-type estimé. Des facteurs de correction fournis par Burr (1969) sont appliqués.

alpha : activez cette option pour définir l'intervalle de confiance autour de la ligne centrale de la carte de contrôle. $(100 - \alpha)\%$ de la distribution se trouve dans les limites de contrôle. Des facteurs de correction fournis par Burr (1969) sont appliqués.

Moyenne : activez cette option pour entrer la valeur de la ligne centrale de la carte de contrôle. Cette valeur est en général calculée à partir d'un historique.

Sigma : activez cette option pour entrer la valeur de l'écart-type du procédé. Cette valeur est en général calculée à partir d'un historique. Si cette option est activée, vous ne pouvez pas choisir une méthode d'estimation de sigma dans l'onglet "Estimation".

Onglet **Estimation** :

Méthode pour Sigma : choisissez la méthode utilisée pour l'estimation de l'écart-type de la carte de contrôle (voir la section description pour plus de détails) :

- Ecart-type global
- R-barre
- S-barre
- Etendue mobile moyenne
- Etendue mobile médiane
- **Longueur des EM** : changez cette valeur pour modifier le nombre de valeurs prises en compte pour le calcul des étendues mobiles.
- Ecart-type

Onglet **Plan** :

Cet onglet n'est actif que si l'option carte CUMSUM est sélectionée

Schéma : Choisissez une des options suivantes en fonction du type de carte souhaité (voir description pour plus de details) :

- **Unilatéral (LCL/UCL)** : les sommes cumulées des bornes inférieures et supérieures sont calculées séparément pour chaque point.

- **FIR** : Activez cette option pour modifier les valeurs initiales des sommes cumulées supérieures et inférieures.
- **Bilatéral (Masque U)** : Les valeurs normalisées sont affichées. A partir du point d'origine, les limites supérieure et inférieure de la détection du décalage moyen sont affichées à l'envers sous la forme d'un masque.
- **Origine** : Activez cette option pour modifier l'origine du masque. La valeur par défaut est le dernier point des données.

Design : Ici vous pouvez entrer les paramètres du décalage moyen (voir description pour plus de details) :

- **h** : Entrez le seuil pour les sommes cumulées supérieures et inférieures ou le masque au-dessus duquel un décalage moyen est détecté.
- **k** : Entrez la granularité de la détection du décalage moyen. k est la moitié du décalage moyen à détecter. la valeur par défaut est 0.5 pour détecter un décalage moyen de 1 sigma.

Onglet **Sorties** :

Afficher les zones : activez cette option pour afficher en plus des limites de contrôle, les limites des zones A et B.

Tests de normalité : activez cette option pour tester la normalité des données (voir l'outil Tests de normalité pour plus de détails).

Niveau de signification (%) : entrez le niveau de signification à utiliser pour les tests (valeur par défaut : 5%).

Test des causes spéciales : activez cette option pour analyser les points de la carte de contrôle en utilisant les règles pour les causes spéciales. Vous pouvez utiliser les règles suivantes :

- 1 point au-delà de 3s de la ligne centrale
- 9 points consécutifs du même côté de la ligne centrale
- 6 points consécutifs tous montant ou tous descendant
- 14 points consécutifs alternant au-dessus et en-dessous
- 2 sur 3 points > 2s de la ligne centrale (du même côté)
- 4 sur 5 points > 1s de la ligne centrale (du même côté)
- 15 points consécutifs plus proche que 1s de la ligne centrale (des deux côtés)
- 8 points > 1s de la ligne centrale (des deux côtés)
- **Toutes** : cliquer sur ce bouton pour sélectionner toutes les options.

- **None** : cliquer sur ce bouton pour désélectionner toutes les options.

Appliquer les règles de Westgard : activez cette option pour analyser les différents points de la carte de contrôle en utilisant les règles de Westgard. Vous pouvez choisir parmi les règles suivantes :

- Règle 1 2s
- Règle 1 3
- Règle 2 2s
- Règle 4s
- Règle 4 1s
- Règle 10 X
- **Toutes** : cliquer sur ce bouton pour sélectionner toutes les options.
- **None** : cliquer sur ce bouton pour désélectionner toutes les options.

Onglet **Graphiques** :

Afficher les graphiques : activez cette option pour visualiser les cartes de contrôle sous forme de graphiques.

Graphiques Q-Q (loi normale) : activez cette option pour afficher des graphiques Q-Q basés sur la loi normale.

Histogrammes : activez cette option pour afficher les histogrammes des échantillons. Pour la distribution théorique, la fonction de densité est affichée.

Run chart : activez cette option pour afficher un graphique figurant les observations de chacun des sous groupes.

Résultats

Estimation :

Moyenne estimée : dans ce tableau sont affichées les moyennes estimées pour les différentes phases.

Ecart-type estimé : dans ce tableau sont affichés les écarts-types estimés pour les différentes phases.

Transformation Box-Cox :

Lambda : ce tableau n'est affiché que si l'option d'optimisation de Lambda a été choisie.

Série avant et après transformation : dans ce tableau sont affichées la série avant transformation et la série après transformation. Si Lambda a été optimisé, la série après optimisation correspond aux résidus du modèle. Si Lambda est fixé, la série après transformation correspond à l'application directe de la transformation de Box-Cox.

Capacités du processus :

Capacités du processus : ces tableaux sont affichés si l'option "Capacités des processus" est activée. Il y a un tableau par phase. Un tableau comprend les indicateurs de capacité du processus et si possible les intervalles de confiance correspondant : Cp, Cpl, Cpu, Cpk, Pp, Ppl, Ppu, Ppk, Cpm, Cpm (Boyle), Cp 5.5, Cpk 5.5, Cpmk, and Cs (Wright).

Pour les Cp, Cpl, et Cpu, une information concernant la performance du procédé est fournie, et pour le Cp une information sur la situation est donnée pour faciliter l'interprétation.

Aux valeurs de Cp values sont associées les états suivants selon Ekvall et Juran (1974):

- "pas adéquat" si Cp < 1
- "adéquat" si 1 <= Cp <= 1.33
- "plus qu'adéquat" si Cp > 1.33

D'après Montgomery (2001), le Cp doit avoir les valeurs minimales suivantes pour que la performance du procédé soit comme attendue :

- 1.33 pour les procédés existants
- 1.50 pour de nouveaux procédés ou des procédés existants si la variable est critique
- 1.67 pour de nouveaux procédés si la variable est critique

D'après Montgomery (2001), le Cpu et le Cpl doivent avoir les valeurs minimales suivantes pour que la performance du procédé soit comme attendue :

- 1.25 pour les procédés existants
- 1.45 pour de nouveaux procédés ou des procédés existants si la variable est critique
- 1.60 pour de nouveaux procédés si la variable est critique

Capacités : ce graphique présente l'information concernant les spécifications et les limites de contrôle. La ligne qui joint les limites inférieures et supérieures correspond aux limites inférieures et supérieures, tandis que le barre verticale correspond à la ligne centrale. Les limites de contrôle correspondant aux différentes phases sont affichées séparément.

Graphiques :

Les éléments suivants sont affichés pour chaque graphique sélectionné. Les résultats ci-dessous sont affichés séparément pour chaque carte demandée. Une carte peut être choisie

seule ou en combinaison avec la carte X individuelle.

Carte UWMA / EWMA / CUSUM : dans ce tableau sont contenues les informations concernant la ligne centrale et les limites de contrôle du graphique en question. Une colonne est affichée pour chaque phase.

Détails pour les observations : dans ce tableau sont affichées des informations détaillées pour chaque sous-groupe. Pour chaque sous-groupe, sont affichés, la phase correspondante, l'effectif, la moyenne, les valeurs minimales et maximales, la ligne centrale et les limites de contrôle inférieure et supérieure. Si l'information concernant les zones A, B et C est activée, alors les limites de contrôle inférieure et supérieure des zones A et B sont aussi affichées.

Détails pour les règles : si l'option pour l'application des règles est active, un tableau détaillé concernant les règles est affiché. Pour chaque sous-groupe, il y a une ligne pour chaque règle à appliquer. "Oui" indique que la règle en question a été appliquée, et "Non" indique que la règle ne s'applique pas.

Carte UWMA / EWMA / CUSUM : si l'option d'affichage des graphiques est active, alors un graphique construit à partir des tableaux mentionnés ci-dessus est affiché. La ligne centrale ainsi que les limites de contrôle inférieures et supérieures sont aussi affichées. Si les options correspondantes ont été activées, les limites de contrôle des zones A et B sont incluses, et des libellés sont affichés pour les sous-groupes pour lesquels les règles sont appliquées. Une légende comprenant les règles appliquées est affichée sous le graphique.

Tests de normalité :

Pour chaque test demandé sont affichées les statistiques relatives au test, dont notamment la p-value qui est ensuite utilisée pour l'interprétation du test par comparaison avec le seuil de signification choisi.

S'ils ont été demandés, les graphiques P-P et Q-Q sont ensuite affichés.

Histogrammes : les histogrammes sont affichés. Si vous le souhaitez, vous pouvez modifier la couleur des lignes, les échelles, et les titres comme avec n'importe quel graphique Excel.

Run chart : le graphique des derniers points est affiché.

Exemple

Un tutoriel expliquant comment utiliser les cartes de contrôle pondérées par le temps est disponible sur le Centre d'aide XLSTAT sur :

http://www.xlstat.com/demoSPW_FR.htm

Bibliographie

Burr, I. W. (1967). The effect of non-normality on constants for X and R charts. *Industrial Quality control*, **23(11)**, 563-569.

Burr I. W. (1969). Control charts for measurements with varying sample sizes. *Journal of Quality Technology*, **1(3)**, 163-167.

Deming W. E. (1993). The New Economics for Industry, Government, and Education. Cambridge, MA: Center for Advanced Engineering Study, Massachusetts Institute of Technology.

Ekvall D. N. (1974). Manufacturing Planning. In Quality Control Handbook,. 3rd Ed. (J. M. **Juran**, et al. eds.) pp. 9-22-39, McGraw-Hill Book Co., New York.

Montgomery D.C. (2001), Introduction to Statistical Quality Control, 4th edition, John Wiley & Sons.

Nelson L.S. (1984). The Shewhart Control Chart - Tests for Special Causes. *Journal of Quality Technology*, **16**, 237-239.

Pyzdek Th. (2003). The Six Sigma Handbook Revised and Expanded, McGraw Hill, New York.

Ryan Th. P. (2000). Statistical Methods for Quality Improvement, Second Edition, Wiley Series in probability and statistics, John Wiley & Sons, New York.

Shewhart W. A. (1931). Economic Control of Quality of Manufactured Product, Van Nostrand, New York.

Diagrammes de Pareto

Utilisez cet outil pour calculer des statistiques descriptives et pour afficher des diagrammes de Pareto pour un ensemble de variables qualitatives.

Dans cette section :

Description

Le diagramme de Pareto doit son nom à l'économiste italien Pareto, mais c'est J. M. Juran qui est considéré comme étant le premier à l'avoir utilisé dans le domaine industriel.

Les causes qui doivent être investiguées (responsables des défauts) sont listées et leurs pourcentages respectifs sont calculés. Les pourcentages sont ensuite utilisés pour construire un diagramme en bâtons.

Vous pouvez sélectionner plusieurs échantillons, mais XLSTAT fait les calculs pour chaque échantillon indépendamment. Un graphique permettant de comparer les échantillons peut être affiché.

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.

OK : cliquez sur ce bouton pour lancer les calculs.

Annuler : cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

Aide : cliquez sur ce bouton pour afficher l'aide.

 : cliquez sur ce bouton pour rétablir les options par défaut.

 : cliquez sur ce bouton pour effacer les sélections de données.

 : cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.

Onglet **Général**:

Causes : sélectionnez une colonne (ou une ligne en mode ligne) de données qualitatives qui représentent la liste des causes pour lesquelles vous voulez calculer les statistiques descriptives.

Effectifs : activez cette option, si vos données sont déjà agrégées en une liste de causes et en la liste correspondante donnant les effectifs de ces causes. Sélectionnez alors ici la colonne contenant les effectifs.

Sous-échantillons : activez cette option pour sélectionner une colonne indiquant les noms ou les indices des sous-échantillons (ou groupes) correspondant à chacune des observations.

- **Libellés variable-modalité** : activez cette option pour que les occurrences des sous-échantillons soit affichées avec le nom de la variable de sous-échantillon suivi de la modalité.

- **Comparer à l'échantillon total** : activez cette option pour que les statistiques descriptives et les graphiques soient aussi affichés pour l'échantillon total.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des variables : activez cette option si la première ligne des sélections contient un libellé.

Poids : activez cette option si vous voulez pondérer les observations. Si vous n'activez pas cette option, les poids seront tous considérés comme valant 1. Les poids doivent être impérativement supérieurs ou égaux à 0. Si un en-tête de colonne a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée. Cette option n'est disponible que si l'option "Effectifs" n'est pas activée.

Onglet **Options**:

Options de tri : choisissez la méthode de tri à appliquer aux causes.

- **Pas de tri** : aucun tri n'est appliqué,
- **Décroissant** : activez cette option pour que, pour chaque modalité, le tri soit décroissant en fonction des effectifs.
- **Première décroissante** : ce tri ne sera différent du précédent que si vous avez sélectionné plusieurs colonnes de causes ou d'effectifs. Le tri descendant est appliqué

pour la première série de causes. Ensuite l'ordre appliqué à la première série de causes est respecté pour les séries suivantes.
- **Alphabétique** : choisissez cette option pour que le tri alphabétique soit appliqué à chaque série de causes.

Regroupement des causes : activez cette option si vous souhaitez regrouper des causes.

- **Effectif inférieur à** : choisissez cette option pour regrouper les modalités dont l'effectif est inférieur à la valeur entrée.
- **% inférieur à** : choisissez cette option pour regrouper les modalités représentant un % de l'échantillon inférieur à la valeur entrée.
- **K effectifs les plus faibles** : choisissez cette option pour regrouper les k modalités présentant les effectifs les plus faibles. k est défini par l'utilisateur.
- **% cumulé** : choisissez cette option pour regrouper les modalités dès lors que le % d'observations cumulé sur les modalités précédentes dépasse la valeur entrée.

Onglet **Données manquantes** :

Ne pas accepter les valeurs manquantes : activez cette option pour que XLSTAT empêche la poursuite des calculs si des valeurs manquantes sont détectées.

Supprimer les observations : activez cette option pour supprimer les observations contenant des données manquantes.

Onglet **Graphiques** :

Axe gauche : choisissez les données à afficher sur l'axe des ordonnées à gauche :

- **Effectifs** : choisissez cette option pour que l'échelle des graphiques corresponde aux effectifs des modalités.
- **Fréquences** : choisissez cette option pour que l'échelle des graphiques corresponde aux fréquences des modalités.

Changer la couleur à : choisissez à partir de quel % cumulé les bâtons du diagramme changent de couleur.

% cumulé : activez cette option pour que la ligne des % cumulés soit affichée.

Toutes les séries sur un graphique : activez cette option pour que, si vous avez sélectionné plusieurs séries de causes ou d'effectifs, vous puissiez les comparer sur un même graphique.

Exemple

Un tutoriel expliquant comment créer un diagramme de Pareto est disponible sur le Centre d'aide XLSTAT :

http://www.xlstat.com/demo-pto.htm

Bibliographie

Juran J.M. (1960). Pareto, Lorenz, Cournot, Bernouli, Juran and others. *Industrial Quality-Control*, **17(4)**, 25.

Pareto V. (1906). Manuel d'Economie Politique. 1. Edition, Paris.

Pyzdek Th. (2003). The Six Sigma Handbook Revised and Expanded, McGraw Hill, New York.

Ryan Th. P. (2000). Statistical Methods for Quality Improvement, Second Edition, Wiley Series in probability and statistics, John Wiley & Sons, New York.

Gage R&R pour variables quantitatives (Analyse du système de mesures)

Utilisez cet outil pour contrôler et valider votre système de mesure, si vous disposez de plusieurs mesures quantitatives prises par un ou plusieurs opérateurs.

Dans cette section :

Description

L'analyse de système de mesure (Measurement System Analysis ou MSA en anglais) aussi désignée par Gage R&R (Gage Repeatability and Reproducibility) est une méthodologie qui permet de contrôler et de valider un processus de mesure. Elle permet notamment d'identifier quelles sont les sources responsables de la variabilité des mesures effectuées : la variabilité peut être due au système de mesure, à l'opérateur effectuant la mesure ou aux objets mesurés. La méthodologie Gage R&R appliquée aux mesures quantitatives s'appuie sur deux outils communs de l'analyse statistique : l'ANOVA et les cartes de contrôle R.

Le mot "gage" (signifiant jauge) fait référence au fait que la méthodologie a été développée pour valider des instruments de mesure.

Une mesure est "répétable" si les mesures obtenues par un opérateur donné pour un même objet (produit, unité, pièce ou échantillon, en fonction du domaine d'application) à plusieurs reprises, ne varient pas au-delà d'un seuil donné. Si la répétabilité d'un système de mesure n'est pas satisfaisante, il convient de s'interroger sur la qualité du système de mesure, ou de former les opérateurs qui n'obtiennent pas de résultats stables.

Une mesure est "reproductible" si les mesures obtenues pour un objet donné (produit, unité, pièce ou échantillon, en fonction du domaine d'application) par plusieurs opérateurs ne varient pas au-delà d'un seuil donné. Si la reproductibilité d'un système de mesure n'est pas satisfaisante, il faut former les opérateurs de telle sorte que leurs résultats soient plus homogènes.

L'objectif d'une analyse Gage R&R est d'identifier les sources de variabilité et de prendre les mesures nécessaires pour les réduire si besoin.

Lorsque les mesures sont des données quantitatives, deux méthodes sont disponibles pour l'analyse Gage R&R. La première est basée sur l'analyse de variance (ANOVA) et la seconde sur les cartes de contrôle R (Amplitude et moyenne).

Dans les descriptions ci-dessous, $\hat{\sigma}^2_{Repeatability}$ désigne la variance liée à la répétabilité. Plus elle est faible, plus la mesure est répétable (les opérateurs donnent chacun des résultats cohérents pour une même pièce). Le calcul de cette variance est différent en fonction de la méthode choisie (ANOVA ou carte de contrôle R). $\hat{\sigma}^2_{\text{Reproducibility}}$ désigne la part de la variance correspondant à la reproductibilité. Plus elle est faible, plus la mesure est reproductible (les divers opérateurs donnent des résultats concordants pour une même pièce). Le calcul de cette variance dépend aussi de la méthode choisie (ANOVA ou carte de contrôle R)

$\hat{\sigma}^2_{R\&R}$ est la variance du gage R&R. Elle est toujours la somme des deux variances précédentes : $\hat{\sigma}^2_{R\&R} = \hat{\sigma}^2_{Repeatability} + \hat{\sigma}^2_{\text{Reproducibility}}$.

ANOVA :

Lorsque le modèle ANOVA est utilisé dans l'analyse R&R, on peut tester statistiquement si la variabilité des mesures est liée aux opérateurs et/ou aux pièces elles-mêmes et/ou à une interaction entre les deux (certains opérateurs pourraient donner pour certaines pièces des mesures sensiblement supérieures ou inférieures). Deux plans expérimentaux sont disponibles pour une analyse Gage R&R : le plan croisé (équilibré) et le plan imbriqué.

Plan croisé :

Une ANOVA équilibrée à deux facteurs (Opérateur et Pièce) est effectuée. On peut choisir le modèle réduit comprenant uniquement les effets principaux, ou le modèle complet qui inclut l'interaction (Pièce*Opérateur). Pour une ANOVA croisée, les données doivent correspondre aux exigences d'un plan équilibré. Cela signifie que chaque pièce doit avoir été mesurée par chacun des opérateurs un même nombre de fois

Pour le modèle complet, les statistiques de Fisher F sont calculées comme suit :

$$F_{Opérateur} = MSE_{Opérateur} / MSE_{Pièce*Opérateur}$$

$$F_{Pièce} = MSE_{Pièce} / MSE_{Pièce*Opérateur}$$

avec MSE désignant la moyenne des carrés des erreurs.

Si la p-Value associée à l'interaction Opérateur*Pièce est plus grande ou égale à une valeur seuil fixée par l'utilisateur (typiquement 25 %), le terme d'interaction est supprimé du modèle et le modèle réduit est alors recalculé.

Dans le cas d'un modèle croisé avec interaction, les variances sont définies comme suit :

$$\hat{\sigma}^2 = MSE_{Error}$$
$$\hat{\sigma}^2_{Pièce*Opérateur} = \left(MSE_{Pièce*Opérateur} - MSE_{Error}\right)/nRep$$
$$\hat{\sigma}^2_{Opérateur} = \left(MSE_{Opérateur} - MSE_{Pièce*Opérateur}\right)/(nPièce.nRep)$$
$$\hat{\sigma}^2_{Pièce} = \left(MSE_{Pièce} - MSE_{Pièce*Opérateur}\right)/(nOpérateur.nRep)$$
$$\hat{\sigma}^2_{Repeatability} = \hat{\sigma}^2$$
$$\hat{\sigma}^2_{\text{Reproducibility}} = \hat{\sigma}^2_{Opérateur} + \hat{\sigma}^2_{Pièce*Opérateur}$$
$$\hat{\sigma}^2_{R\&R} = \hat{\sigma}^2_{Repeatability} + \hat{\sigma}^2_{\text{Reproducibility}}$$

Dans le cas d'un modèle réduit (sans interaction), les variances sont définies comme suit :

$$\hat{\sigma}^2 = MSE_{Error}$$
$$\hat{\sigma}^2_{Pièce*Opérateur} = 0$$
$$\hat{\sigma}^2_{Opérateur} = \left(MSE_{Opérateur}\right)/(nPièce.nRep)$$
$$\hat{\sigma}^2_{Pièce} = \left(MSE_{Pièce}\right)/(nOpérateur.nRep)$$
$$\hat{\sigma}^2_{Repeatability} = \hat{\sigma}^2$$
$$\hat{\sigma}^2_{\text{Reproducibility}} = \hat{\sigma}^2_{Opérateur} + \hat{\sigma}^2_{Pièce*Opérateur}$$
$$\hat{\sigma}^2_{R\&R} = \hat{\sigma}^2_{Repeatability} + \hat{\sigma}^2_{\text{Reproducibility}}$$

avec MSE désignant la moyenne des carrés des erreurs, $nRep$ le nombre de répétitions, $nPièce$ le nombre de pièces et $nOpérateur$ le nombre d'opérateurs.

Plan imbriqué :

Dans ce cas une ANOVA à deux facteurs imbriqués est calculée avec les facteurs Operateur et Pièce(Opérateur).

Les conditions suivantes sont requises : la fréquence d'occurrence doit être la même pour toutes les catégories, et chaque pièce ne doit être inspectée que par un seul opérateur. Les statistiques F sont calculées comme suit :

$$F_{Opérateur} = MSE_{Opérateur}/MSE_{Pièce(Opérateur)}$$

$$F_{Pièce(Opérateur)} = MSE_{Pièce(Opérateur)}/MSE_{Error}$$

où MSE correspond à la moyenne des carrés des erreurs.

$$\hat{\sigma}^2 = MSE_{Error}$$
$$\hat{\sigma}^2_{Repeatability} = \hat{\sigma}^2$$
$$\hat{\sigma}^2_{\text{Reproducibility}} = \left(MSE_{Opérateur} - MSE_{Pièce(Opérateur)}\right)/(nPièce.nRep)$$
$$\hat{\sigma}^2_{R\&R} = \hat{\sigma}^2_{Repeatability} + \hat{\sigma}^2_{\text{Reproducibility}}$$

où $nRep$ est le nombre de répétitions, $nPièce$ est le nombre de pièces, et $nOpérateur$ est le nombre d'opérateurs.

R charts :

L'analyse Gage R&R fondée sur les cartes de contrôle R, bien que moins puissante que celle s'appuyant sur l'ANOVA, présente l'avantage d'être plus simple à calculer et produit des cartes de contrôle (cartes R). Comme l'analyse de la variance, cette méthode permet de calculer la répétabilité et la reproductibilité du système de mesure. Pour utiliser cette méthode, vous avez besoin d'avoir plusieurs pièces, des opérateurs et des répétitions (typiquement 10 pièces, 3 opérateurs et 2 répétitions).

Avec cette approche, les diverses variances sont calculées comme suit :

$$\hat{\sigma}^2_{Repeatability} = \bar{R}/d_2^*(nRep, nPièce * nOpérateur)$$
$$\hat{\sigma}^2_{\text{Reproducibility}} = \left(\frac{\text{Max}(\mu_{Pièce}) - \text{Min}(\mu_{Pièce})}{d_2^*(nOpérateur, 1)}\right)^2 - \frac{\hat{\sigma}^2_{Repeatability}}{nPièce * nOpérateur}$$
$$\hat{\sigma}^2_{R\&R} = \hat{\sigma}^2_{Repeatability} + \hat{\sigma}^2_{\text{Reproducibility}}$$
$$\hat{\sigma}^2_{Pièce} = \left(\frac{\text{Max}(\mu_{Opérateur}) - \text{Min}(\mu_{Opérateur})}{d_2^*(nPièce, 1)}\right)^2$$
$$\hat{\sigma}^2 = \hat{\sigma}^2_{R\&R} + \hat{\sigma}^2_{Pièce}$$

où $\text{Max}(\mu_{Pièce}\text{ (respectivement Opérateur)}) - \text{Min}(\mu_{Pièce}\text{ (respectivement Opérateur)})$ est la différence entre le maximum et le minimum inter-opérateurs (respectivement pièces) des moyennes pour chacune des pièces (respectivement opérateurs), $nRep$ est le nombre de répétitions, $nPièce$ est le nombre de pièces, $nOpérateur$ est le nombre d'opérateurs et $d_2^*(m, k)$ est la constante de la carte de contrôle telle que définie par Burr (1969).

Pendant le calcul de la répétabilité, on voit que la moyenne de l'amplitude de la carte R est utilisée. La variabilité des pièces et la reproductibilité sont fondées sur les valeurs moyennes de la carte $\bar{X}$.

Indicateurs :

XLSTAT propose divers indicateurs construits à partir des variances pour décrire le système de mesure.

La variation de l'étude (*Study variation* en anglais) pour les différentes sources est calculée comme le produit de l'écart-type correspondant par le facteur k Sigma :

$$Variation\ de\ l'étude = k * \hat{\sigma}$$

La tolérance exprimée en pourcentage est définie comme le rapport de la variation de l'étude et la tolérance définie par l'utilisateur :

$$\%\ tolérance = Variation\ de\ l'étude/tolérance$$

Le sigma du processus exprimé en pourcentage est défini comme le rapport de l'écart-type de la source et du sigma historique du processus :

$$\%processus = écart - type\ de\ la\ source/sigma\ processus$$

Rapport Précision sur Tolérance ratio (P/T) :

$$P/T = \frac{k * \hat{\sigma}^2_{R\&R}}{tolerance}$$

Rho P (Rho Pièce) :

$$\rho_{Pièce} = \frac{\hat{\sigma}^2_{Pièce}}{\hat{\sigma}^2}$$

Rho M :

$$\rho_M = \frac{\hat{\sigma}^2_{R\&R}}{\hat{\sigma}^2}$$

Rapport signal bruit (*Signal to Noise Ratio* SNR) :

$$SNR = \sqrt{\frac{2\rho_{Pièce}}{1 - \rho_{Pièce}}}$$

Rapport de discrimination (*Discrimination Ratio* DR) :

$$DR = \frac{1 + \rho_{Pièce}}{1 - \rho_{Pièce}}$$

Biais :

$$Biais = \mu_{Measurements} - \text{cible}$$

Biais en pourcentage :

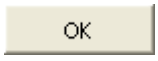
$$Biais\% = (\mu_{Measurements} - cible)/tolérance$$

Résolution :

$$Résolution = Biais + 3 * \hat{\sigma}^2_{R\&R}$$

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

Annuler : cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

Aide : cliquez sur ce bouton pour afficher l'aide.

 : cliquez sur ce bouton pour rétablir les options par défaut.

 : cliquez sur ce bouton pour effacer les sélections de données.

 : cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.

 : cliquez sur ce bouton pour changer la façon dont vous chargez les données dans XLSTAT. Si le bouton possède une icône de souris, XLSTAT vous permet de faire une sélection à la souris des données. Si le bouton possède une icône de liste, XLSTAT vous permet de sélectionner les données à partir d'une liste. Si le bouton possède une icône orange, des boutons complémentaires avec un point d'interrogation  sont affichés pour vous permettre d'importer les données à partir de fichiers.

Onglet **Général** :

Y / Mesures : sélectionnez la colonne ou la ligne qui contient toutes les données de mesure.

X / Opérateurs : sélectionnez la colonne ou la ligne qui contient toutes les données indiquant quel opérateur a effectué chaque mesure.

Pièces : sélectionnez la colonne ou la ligne qui contient toutes les données indiquant sur quelle pièce a été effectuée chaque mesure.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des colonnes/lignes : activez cette option si la première ligne (mode colonnes) ou colonne (mode lignes) des données sélectionnées contient des libellés.

Tri alphabétique des modalités : activez cette option pour que dans les divers résultats, les modalités soient triées alphabétiquement pour la variable d'opérateurs et la variable de pièces.

Onglet **Options/Model** :

Méthode : choisissez la méthode que vous voulez utiliser :

- **ANOVA** : activez cette option, pour utiliser le modèle d'ANOVA pour l'analyse R&R.
- **R chart** : activez cette option, pour utiliser le modèle de carte de contrôle R pour l'analyse R&R.

k Sigma : entrez la valeur de la dispersion. La valeur par défaut est 6.

Intervalle de tolérance : activez cette option pour définir l'amplitude de l'intervalle de tolérance $(USL - LSL)$.

Sigma du procédé : activez cette option pour entrer la valeur du sigma utilisée pour la carte de contrôle. Cette valeur doit être basée sur des données historiques.

Cible : activez cette option pour entrer la valeur de référence pour les mesures.

ANOVA : dans le cas ou la méthode ANOVA a été choisie, sélectionnez le modèle d'ANOVA qui doit être utilisé lors de l'analyse :

- Réduit
- Croisé
- **Niveau de signification (%)** : entrez le niveau de signification pour le test F qui permet de déterminer si l'interaction doit être prise en compte (valeur par défaut : 25%).
- Imbriqué

Onglet **Options/Estimation** :

Méthode pour Sigma : choisissez la méthode à utiliser pour l'estimation de l'écart-type de la carte de contrôle (voir la section description pour plus de détails) :

- Ecart-type total
- R-bar
- S-bar

Onglet **Sorties** :

Composantes de la variance : activez cette option pour afficher le tableau présentant les différentes composantes de la variance.

Indicateur d'état : activez cette option pour afficher les indicateurs d'état pour l'analyse du système de mesure.

Analyse de la variance : activez cette option pour afficher le tableau d'analyse de la variance.

Onglet **Graphiques** :

Afficher les graphiques : activez cette option pour visualiser les cartes de contrôle.

- **Afficher les zones** : activez cette option pour afficher en plus des limites de contrôle, les limites des zones B et C.

Box plots : activez cette option pour afficher les box plots. Voir la section description des graphiques univariés pour plus de détails.

Scattergrams : activez cette option pour afficher les scattergrams. La moyenne (croix + rouge) et la médiane (ligne rouge) sont toujours affichées.

- **Minimum/Maximum** : activez cette option pour systématiquement afficher les points correspondant au minimum et au maximum (box plots).
- **Valeurs extrêmes** : activez cette option pour afficher les points correspondant aux valeurs extrêmes (box plots) avec un cercle évidé.
- **Position des étiquettes** : choisissez la position des étiquettes sur les graphiques verticaux. Elles peuvent être soit en bas, soit en haut, soit alternativement en bas et en haut.

Graphiques des moyennes : activez cette option pour afficher les graphiques permettant d'afficher les moyennes des différentes modalités des facteurs.

Résultats

Composantes de la variance :

Le premier tableau et le graphique associé présentent les différentes sources de variabilité. Les contributions à la variance totale et à la variance de l'étude, calculées en utilisant la valeur de dispersion fournie par l'utilisateur, sont ensuite affichées.

Si un intervalle de tolérance a été défini, la distribution de la variabilité en fonction de l'intervalle de tolérance fournie est affichée.

Si le sigma du processus a été fourni, la distribution de la variabilité en fonction du sigma du processus est affichée.

Le tableau suivant présente la répartition détaillée de la variance en fonction des différentes sources. Les valeurs absolues des composantes de variance et le pourcentage de la variance totale sont affichés.

Le troisième tableau montre la répartition de l'écart-type pour les différentes sources. Il affiche les valeurs absolues des composantes de la variance, la variation de l'étude qui est calculée

comme le produit de l'écart-type et de la dispersion, le pourcentage de la variation de l'étude, la variabilité de la tolérance, définie comme le rapport entre la variation de l'étude et sigma du processus, et le pourcentage de la variabilité du processus.

Indicateurs de statut :

Le premier tableau montre les informations relatives à l'évaluation du système de mesure. Le rapport précision sur tolérance (P/T), le Rho P, le Rho M, le rapport signal bruit (SNR), le rapport de discrimination ratio (DR), le biais absolu et relatif (%) et la résolution sont affichés.

Les valeurs P/T sont interprétées comme suit :

- "plus qu'adéquat" si P/T <= 0.1
- "adéquat" si 0.1 < P/T <= 0.3
- "pas adéquat" si P/T > 0.3

Les valeurs du SNR sont interprétées comme suit :

- "pas acceptable" si SNR < 2
- "pas adéquat" si 2 <= SNR <= 5
- "adéquat" si SNR > 5

Coefficients d'ajustement : dans ce tableau sont affichées les statistiques relatives à l'ajustement du modèle d'ANOVA :

- **Observations** : le nombre d'observations prises en compte dans les calculs. Dans les formules présentées ci-dessous n désigne le nombre d'observations.
- **Somme des poids** : la somme des poids des observations prises en compte dans les calculs. Dans les formules présentées ci-dessous W désigne la somme des poids.
- **DDL** : le nombre de degrés de liberté pour le modèle retenu (correspondant à la partie erreurs).
- **R²** : le coefficient de détermination du modèle. Ce coefficient, dont la valeur est comprise entre 0 et 1, n'est affiché que si la constante du modèle n'est pas fixée par l'utilisateur. Sa valeur est définie par

$$R^2 = 1 - \frac{\sum_{i=1}^{n} w_i (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} w_i (y_i - \bar{y})^2} \text{ avec } \bar{y} = \frac{1}{n} \sum_{i=1}^{n} w_i y_i$$

- Le R^2 s'interprète comme la proportion de la variabilité de la variable dépendante expliquée par le modèle. Plus le R^2 est proche de 1, meilleur est le modèle. L'inconvénient du R^2 est qu'il ne prend pas en compte le nombre de variables utilisées pour ajuster le modèle.

- **R²ajusté** : le coefficient de détermination ajusté du modèle. Le R^2 ajusté peut être négatif si le R^2 est voisin de zéro. Ce coefficient n'est affiché que si la constante du modèle n'est pas fixée par l'utilisateur. Sa valeur est définie par

$$\hat{R}^2 = 1 - (1 - R^2)\frac{W - 1}{W - p - 1}$$

Le R² ajusté est une correction du R² qui permet de prendre en compte le nombre de variables utilisées dans le modèle.

- **MCE** : la moyenne des carrés des erreurs (MCE) est définie par :

$$MCE = \frac{1}{W - p^*}\sum_{i=1}^{n} w_i(y_i - \hat{y}_i)^2$$

- **RMCE** : la racine de la moyenne des carrés des erreurs (RMCE) est la racine carrée de la MCE.

Analyse de la variance :

Le **tableau d'analyse de la variance** permet d'évaluer le pouvoir explicatif des variables explicatives.

Graphiques :

Les résultats suivants sont affichés séparément pour chaque graphique requis. Les graphiques.

Carte X barre/ R : dans ce tableau sont contenues les informations concernant la ligne centrale et les limites de contrôle du graphique en question.

Détails pour les observations : dans ce tableau sont affichées des informations détaillées pour chaque sous-groupe (sous-groupe désignant ici un couple Opérateur-Pièce). Pour chaque sous-groupe, sont affichés, la phase correspondante, l'effectif, la moyenne, les valeurs minimales et maximales, la ligne centrale et les limites de contrôle inférieure et supérieure. Si l'information concernant les zones B et C est activée, alors les limites de contrôle inférieures et supérieures des zones B and C sont aussi affichées.

Carte X barre / carte R : si l'option d'affichage des graphiques est activée, alors un graphique construit à partir des tableaux mentionnés ci- dessus est affiché. Chaque sous-groupe est affiché. La ligne centrale ainsi que les limites de contrôle inférieures et supérieures sont aussi affichées. Si les options correspondantes ont été activées, les limites de contrôle des zones B et C sont incluses, et des libellés sont affichés pour les sous- groupes pour lesquels les règles sont appliquées. Une légende comprenant les règles appliquées est affichée sous le graphique.

Enfin, sont affichés les graphiques présentant les moyennes pour chaque opérateur, pièce et interaction.

Exemple

Un tutoriel expliquant comment utiliser l'analyse Gage R&R est disponible sur le Centre d'aide XLSTAT :

http://www.xlstat.com/demo-rrxf.htm

Bibliographie

Burr, I. W. (1967). The effect of non-normality on constants for X and R charts. *Industrial Quality control*, **23(11)**, 563-569.

Burr I. W. (1969). Control charts for measurements with varying sample sizes. *Journal of Quality Technology*, **1(3)**, 163-167.

Deming W. E. (1993). The New Economics for Industry, Government, and Education. Cambridge, MA: Center for Advanced Engineering Study, Massachusetts Institute of Technology.

Ekvall D. N. (1974). Manufacturing Planning. In Quality Control Hand-. book,. 3rd Ed. (J. M. **Juran**, et al. eds.) pp. 9-22-39, McGraw-Hill Book Co., New York.

Montgomery D.C. (2001), Introduction to Statistical Quality Control, 4th edition, John Wiley & Sons.

Nelson L.S. (1984). The Shewhart Control Chart - Tests for Special Causes. *Journal of Quality Technology*, **16**, 237-239.

Pyzdek Th. (2003). The Six Sigma Handbook Revised and Expanded, McGraw Hill, New York.

Ryan Th. P. (2000). Statistical Methods for Quality Improvement, Second Edition, Wiley Series in probability and statistics, John Wiley & Sons, New York.

Shewhart W. A. (1931). Economic Control of Quality of Manufactured Product, Van Nostrand, New York.

Gage R&R pour Attributs (Analyse du système de mesures)

Utilisez cet outil pour contrôler et valider une méthode de mesure ou un système de mesure, dans le cas où vous disposez de mesures qualitatives (attributs) nominales ou ordinales relevées par un ou plusieurs opérateurs sur plusieurs pièces.

Dans cette section :

Description

L'analyse de système de mesure (Measurement System Analysis ou MSA en anglais) aussi désignée par Gage R&R (Gage Repeatability and Reproducibility) est une méthodologie qui permet de contrôler et de valider un processus de mesure. Elle permet notamment d'identifier quelles sont les sources responsables de la variabilité des mesures effectuées : la variabilité peut être due au système de mesure, à l'opérateur effectuant la mesure ou aux objets mesurés. Le mot "gage" (signifiant jauge en anglais) fait référence au fait que la méthodologie a été développée pour valider des instruments de mesure.

Les données de type attribut sont des caractéristiques qualitatives (ou attributs) pouvant être catégorisées et comptées. Ces données peuvent être nominales, c'est-à-dire à plusieurs niveaux et sans ordre naturel, ou ordinales, c'est-à-dire avec au moins trois niveaux et un ordre naturel.

Contrairement à la méthodologie Gage R&R pour des mesures quantitatives, l'analyse pour les données qualitatives (attributs) donne des informations sur « l'accord » et la « justesse ». Les notions de variance, de répétabilité et de reproductibilité ne s'appliquent pas dans ce cas.

Un fort « accord » correspond au cas où les mesures prises à plusieurs reprises par un opérateur donné pour le même objet (produit, unité, pièce ou échantillon, en fonction du domaine d'application) sont cohérentes. Si l'accord d'un système de mesure est faible, il convient de s'interroger sur la qualité du système de mesure ou du protocole, ou de former les opérateurs qui ne sont pas performants, si le système de mesure ne semble pas être responsable de l'absence d'accord.

Une bonne « justesse » correspond au cas où les mesures prises par un opérateur pour un même objet à plusieurs reprises (produit, unité, pièce ou échantillon, en fonction du domaine d'application) correspondent aux valeurs données par une méthode ou un expert servant de référence. Si la « justesse » d'un système de mesure est faible, il faut former les opérateurs, afin que leurs résultats soient plus corrects.

Le but d'une analyse Gage R&R pour attributs est d'identifier les sources de faible accord et de faible justesse pour prendre éventuellement les décisions nécessaires.

L'analyse Gage R&R pour attributs est basée sur les statistiques suivantes pour évaluer l'accord et la justesse :

- Statistiques d'accord
- Statistiques Kappa
- Statistiques de Kendall

Si possible, les comparaisons suivantes sont réalisées :

- Intra opérateur
- Inter opérateurs
- Opérateur versus référence
- Tous les opérateurs versus référence

La référence (ou standard) correspond aux mesures rapportées par un expert ou une méthode réputée très fiable.

Statistiques mesurant le degré d'accord

Il est possible de calculer ces statistiques pour tous les types d'évaluation. Pour chaque type d'évaluation, le nombre de pièces évaluées, le nombre d'évaluations concordantes, le ratio d'évaluations concordantes, le pourcentage d'évaluations concordantes et l'intervalle de confiance de ce pourcentage sont affichés.

Dans le cas d'une évaluation intra opérateur, XLSTAT calcule pour l'opérateur le nombre de cas où il donne la même mesure pour les différentes répétitions sur l'ensemble des répétitions.

Dans le cas d'une évaluation inter opérateurs, XLSTAT calcule le nombre de cas où les opérateurs donnent la même mesure sur l'ensemble des répétitions.

Dans le cas d'une évaluation opérateur versus référence, XLSTAT calcule pour l'opérateur le nombre de cas où il donne la même mesure que le standard sur l'ensemble des répétitions.

Dans le cas d'une évaluation pour tous les opérateurs versus référence, XLSTAT calcule le nombre de cas où tous les opérateurs donnent la même mesure que la référence sur l'ensemble des répétitions.

Coefficients Kappa :

Le kappa de Cohen et le kappa de Fleiss sont deux indices adaptés au cas des variables qualitatives. Ces coefficients sont calculés sur des tableaux de contingence provenant de pièces appariées. Le kappa de Fleiss est une généralisation du kappa de Cohen lorsqu'il y a plus de deux juges. Le coefficient kappa varie entre -1 et 1, et permet de mesurer le degré d'accord. Plus il est proche de 1, plus l'accord est important.

Dans le cas d'une évaluation intra opérateur, le kappa de Cohen ne peut être calculé que s'il y a exactement deux répétitions.

Dans le cas d'une évaluation inter opérateurs, le kappa de Cohen ne peut être calculé que pour deux opérateurs avec une évaluation unique.

Coefficients de Kendall :

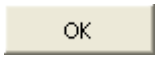
Ces coefficients sont calculés pour des variables ordinales.

Le coefficient de concordance de Kendall permet de mesurer sur une échelle de 0 à 1 le degré de concordance entre deux variables ordinales.

Le coefficient de corrélation de Kendall, aussi appelé tau de Kendall ou tau-b, permet de mesurer sur une échelle allant de -1 à 1 le degré de concordance entre deux variables ordinales.

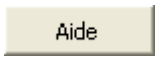
Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.



: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.


: cliquez sur ce bouton pour changer la façon dont vous chargez les données dans XLSTAT. Si le bouton possède une icône de souris, XLSTAT vous permet de faire une sélection à la souris des données. Si le bouton possède une icône de liste, XLSTAT vous permet de sélectionner les données à partir d'une liste. Si le bouton possède une icône orange, des boutons complémentaires avec un point d'interrogation  sont affichés pour vous permettre d'importer les données à partir de fichiers.

Onglet **Général** :

Format de données : préciser le format de données :

- **Tableau observations/variables** : activez cette option si les données contiennent une colonne pour toutes les mesures.
- **Multi-colonne** : activez cette option si les données contiennent une colonne par opérateur et par répétition.

Si le format de données sélectionné est tableau observation/variable :

Mesures : sélectionnez la colonne ou la ligne qui contient toutes les données de mesure.

Préciser le type de données pour les mesures :

- **Ordinales** : activez cette option si les mesures sont ordinales.
- **Nominales** : activez cette option si les mesures sont nominales.

Opérateurs : sélectionnez la colonne ou la ligne qui contient toutes les données indiquant quel opérateur a effectué chaque mesure.

Pièces : sélectionnez la colonne ou la ligne qui contient toutes les données indiquant sur quelle pièce a été effectuée chaque mesure.

Référence : activez cette option si des mesures obtenues par une méthode de référence ou considérées comme standard sont disponibles pour chaque mesure. Sélectionnez la colonne ou la ligne qui contient toutes les données de référence.

Si le format de données sélectionné est multi-colonne :

Mesures : sélectionnez la colonne ou la ligne qui contient toutes les données de mesure.

Préciser le type de données pour les mesures :

- **Ordinales** : activez cette option si les mesures sont ordinales.
- **Nominales** : activez cette option si les mesures sont nominales.

Nombre d'opérateurs : entrez le nombre d'opérateurs ayant évalués les pièces.

Nombre de répétitions : entrez le nombre de répétitions effectuées par chaque opérateur.

Référence : activez cette option si des mesures obtenues par une méthode de référence ou considérées comme standard sont disponibles pour chaque mesure. Sélectionnez la colonne ou la ligne qui contient toutes les données de référence.

Nom des opérateurs : activez cette option si vous connaissez le nom de chaque opérateur. Sélectionnez la colonne ou la ligne qui contient les noms de chaque opérateur.

Plage : activez cette option pour afficher les résultats à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des colonnes/lignes : activez cette option si la première ligne (mode colonnes) ou colonne (mode lignes) des données sélectionnées contient des libellés.

Onglet **Options** :

Statistiques :

- **Kappa de Fleiss** : activez cette option pour calculer le kappa de Fleiss.
- **Kappa de Cohen** : activez cette option pour calculer le kappa de Cohen.
- **Coefficient de concordance de Kendall** : activez cette option pour calculer le coefficient de concordance de Kendall.
- **Coefficient de corrélation de Kendall** : activez cette option pour calculer le coefficient de corrélation de Kendall.
- **Intervalle de confiance (%)** : entrez l'intervalle de confiance (valeur par défaut : 95).

Onglet **Sorties** :

Accord : activez cette option pour afficher le tableau contenant les statistiques d'accord (*agreement statistics*).

Intra opérateur : activez cette option pour afficher les tableaux contenant les résultats intra opérateur.

Entre les opérateurs : activez cette option pour afficher les tableaux contenant les résultats entre les opérateurs.

Opérateur versus la référence : activez cette option pour afficher les tableaux contenant les résultats d'un opérateur comparé à la référence.

Tous les opérateurs versus la référence : activez cette option pour afficher les tableaux contenant les résultats de tous les opérateurs comparé à la référence.

Onglet **Graphiques** :

Graphiques d'accord : activez cette option pour afficher les graphiques qui montrent les moyennes et les intervalles de confiance correspondant pour les statistiques d'accord (agreement).

- **Intra opérateur** : activez cette option pour afficher le graphique d'accord pour chaque opérateur.
- **Opérateur versus la référence** : activez cette option pour afficher le graphique d'accord pour chaque opérateur comparé à la référence.

Résultats

Les résultats sont divisés en quatre sections :

- Intra opérateur
- Inter opérateurs
- Opérateur versus la référence
- Tous les opérateurs versus la référence

A l'intérieur de chaque section, les indicateurs suivants sont affichés, si tant est qu'ils aient été demandés :

- Coefficients d'accord
- Coefficients Kappa
- Coefficients de Kendall

Bibliographie

Agresti A. (1990). Categorical Data Analysis. John Wiley and Sons, New York.

Agresti A., and Coull B.A. (1998). Approximate is better than "exact" for interval estimation of binomial proportions. *The American Statistician*, **52**, 119-126.

Agresti A. and Caffo, B. (2000). Simple and effective confidence intervals for proportions and differences of proportions result from adding two successes and two failures. *The American Statistician*, **54**, 280-288.

Burr, I. W. (1967). The effect of non-normality on constants for X and R charts. *Industrial Quality control*, **23(11)**, 563-569.

Burr I. W. (1969). Control charts for measurements with varying sample sizes. *Journal of Quality Technology*, **1(3)**, 163-167.

Clopper C.J. and Pearson E.S. (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26, 404-413.

Deming W. E. (1993). The New Economics for Industry, Government, and Education. Cambridge, MA: Center for Advanced Engineering Study, Massachusetts Institute of Technology.

Ekvall D. N. (1974). Manufacturing Planning. In Quality Control Hand-. book,. 3rd Ed. (J. M. **Juran**, et al. eds.) pp. 9-22-39, McGraw-Hill Book Co., New York.

Montgomery D.C. (2001), Introduction to Statistical Quality Control, 4th edition, John Wiley & Sons.

Nelson L.S. (1984). The Shewhart Control Chart - Tests for Special Causes. *Journal of Quality Technology*, **16**, 237-239.

Pyzdek Th. (2003). The Six Sigma Handbook Revised and Expanded, McGraw Hill, New York.

Ryan Th. P. (2000). Statistical Methods for Quality Improvement, Second Edition, Wiley Series in probability and statistics, John Wiley & Sons, New York.

Shewhart W. A. (1931). Economic Control of Quality of Manufactured Product, Van Nostrand, New York.

Wilson, E.B. (1927). Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, **22**, 209-212.

Wald, A., & Wolfowitz, J. (1939). Confidence limits for continuous distribution functions. The Annals of Mathematical Statistics, 10, 105-118.

Plans d'expériences

Plans d'effet de facteurs

Utilisez ce module pour générer un plan d'effet de facteurs afin d'identifier l'effet de 2 à 35 facteurs sur une ou plusieurs réponses. Cette famille de plans d'expériences est utilisée afin de trouver les facteurs ayant la plus grande influence parmi tous les facteurs.

Dans cette section :

Description

Les plans d'effet de facteurs sont adaptés à l'étude de l'effet de plusieurs facteurs. Les plans classiquement utilisés dans ce cas sont les plans factoriels. Néanmoins, le nombre d'expériences nécessaire afin d'appliquer un plan factoriel est souvent trop grand dans des cas pratiques. D'autres types de plans sont alors utilisés.

XLSTAT propose de réaliser un plan factoriel complet où l'ensemble des combinaisons de facteurs seront effectuées au cours de l'expérimentation. Ce type de plan est très utiles pour avoir les meilleurs résultats possibles, cependant, le nombre d'expériences peut très vite augmenter. C'est pourquoi XLSTAT propose aussi plusieurs centaines de plans orthogonaux qui sont privilégiés car ils permettent, au moment de l'analyse, d'appliquer une analyse de la variance (ANOVA) sur données équilibrées. Il est alors proposé à l'utilisateur de sélectionner un plan orthogonal proche du plan recherché lors de l'analyse. L'utilisateur peut aussi décider de rechercher un plan optimal. Les plans d'expériences orthogonaux gardés en mémoire incluent la plupart des plans classiques tels que les plans factoriels complets, les plans en carrés latins, les plans de Placket et Burman.

Si aucun des plans orthogonaux présentés n'est adapté à votre analyse, il est alors possible de rechercher un plan D-optimal. Ces plans ne seront pas forcément orthogonaux.

En sortie, afin de faciliter la saisie des réponses, des feuilles individuelles associées à chaque expérience peuvent être générées sur des feuilles Excel séparées, qui pourront être imprimées et remplies.

Modèle

Cet outil génère des plans qui peuvent être analysés en utilisant un modèle additif sans interactions pour l'estimation des effets des facteurs. Si p est le nombre de facteurs, le modèle d'analyse de la variance (ANOVA) est écrit comme suit :

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_{k(i,j),j} + \epsilon_i$$

Plans d'expériences communs

Lorsque l'on débute la création d'un plan d'expériences, la base de données comportant tous les plans d'expérience mémorisés est analysée afin de trouver un plan orthogonal proche des paramètres entrés par l'utilisateur. Une distance entre le plan recherché et le plan issu de la base est alors calculée :

p_i = nombre de facteurs avec i modalités dans le plan recherché

c_i = nombre de facteurs avec i modalités dans le plan issu de la base

p_{exp} = nombre d'expériences dans le plan recherché

c_{exp} = nombre d'expériences dans le plan issu de la base

$d(c,p)$ = $\sum_{i=2}^{7} |c_i - p_i| + c_{exp} - p_{exp}$ (1)

Ainsi, dans l'une des boîtes de dialogue de cet outil, on peut trouver tous les plans ayant le même nombre de facteurs que le plan recherché et tels que $d < 20$.

Le nom formel pour les plans obtenus est donné par :

$$L_n(p_1^{c_1}, ..., p_m^{c_m})$$

Avec :

n = nombre d'expériences

c_i = nombre de modalités du groupe de facteurs p_i

p_i = nombre de facteurs ayant c_i modalités

Le nom utilisé dans la littérature sur les plans d'expérience est aussi donné.

Plan D-Optimal

Pour générer un plan D-Optimal, XLSTAT utilise l'algorithme de Fedorov, qui utilise une méthode de permutation (voir Cook et Nachtsheim, 1980). à chaque itération, un échange

simple est réalisé. L'algorithme va alors échanger le couple qui optimise le plan en fonction du critère décrit ci-après.

La représentation matricielle du plan d'expériences utilise le codage suivant. Pour un facteur f_i ayant c_i modalités, on ajoute $c_i - 1$ colonnes $k_1, ..., k_{c_i - 1}$ à la matrice X du plan de la manière suivante :

f_i		k_{ci-1}	...	k_2	k_1
1		-1		-1	-1
2		0		0	1
3		0		1	0
c_i		1		0	0

La matrice du plan complet X est composée de n lignes, n étant le nombre d'expériences. La matrice comprend une première colonne avec uniquement des 1 et $c_i - 1$ colonnes pour chaque facteur f_i du plan d'expériences, avec c_i le nombre de modalités du facteur f_i.

Le critère utilisé pour l'optimisation est le suivant :

$$c = \log_{10}(\det(X^t X)) \tag{2}$$

Avec :

$X^t X$ = matrice d'information

X = matrice recodée du plan

Ce critère est appelé dans les résultats de la façon suivante :

$$c = \mathrm{Log}(|I|)$$

Le critère suivant est très courant et est également présent dans les résultats :

$$\mathrm{Log}(|I|^{1/p})$$

Lorsqu'on compare des plans d'expérience ayant un nombre d'expériences différent, le logarithme normalisé est utilisé :

$$Norm.\log = \log_{10}((\det \frac{1}{N}(X^t X))^{1/p}) \tag{3}$$

Ce critère est appelé dans les résultats de la façon suivante :

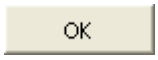
$$Norm.\log = \mathrm{Log}(|1/n * I||^{1/p})$$

Cet indice permet de comparer l'optimalité de différents plans même lorsque le nombre d'expériences varie.

L'utilisateur peut sélectionner un nombre de répétitions afin de trouver un optimum local ayant de bonnes propriétés.

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.


: cliquez sur ce bouton pour changer la façon dont vous chargez les données dans XLSTAT. Si le bouton possède une icône de souris, XLSTAT vous permet de faire une sélection à la souris des données. Si le bouton possède une icône de liste, XLSTAT vous permet de sélectionner les données à partir d'une liste.

Onglet **Général** :

Facteurs quantitatifs (min/max) : sélectionnez le minimum et le maximum des facteurs quantitatifs. La sélection doit comporter deux lignes, la première correspond aux minimums, la seconde aux maximums pour chacun des facteurs.

Facteurs qualitatifs : sélectionnez le tableau des facteurs qualitatifs et leurs modalités.

Nombre de réponses : entrez le nombre de réponses à analyser.

Nombre d'expériences : entrez le nombre d'expériences à mener lors du plan.

Répétitions : activez cette option afin de sélectionner le nombre de répétitions du plan d'expériences.

Ordre aléatoire : activez cette option si vous voulez que l'ordre des expériences dans le plan soit aléatoire.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des variables : activez cette option si la première ligne de la sélection contient le libellé des variables.

Onglet **Options** :

Méthode : Choisir la méthode que vous désirez utiliser pour générer le plan d'expériences.

- **Plan factoriel complet** : sélectionnez cette option afin de réaliser un plan factoriel complet.
- **Plan D-Optimal** : Cette méthode permet de rechercher un plan D-Optimal.
- **Répétitions** : Dans le cas d'une partition initiale aléatoire, entrez le nombre de répétitions à effectuer.
- **Conditions d'arrêt** :
    - **Itérations** : entrez le nombre maximal d'itérations pour l'algorithme. Les calculs sont arrêtés lorsque ce nombre est dépassé. Valeur par défaut: 50.
    - **Convergence** : entrez la valeur maximale pour l'évolution du critère utilisé d'une itération à une autre. Lorsque cette valeur est atteinte, l'algorithme a convergé. Valeur par défaut: 0,00001.
- **Plan orthogonal** : cette méthode permet de chercher un plan orthogonal proche des paramètres entrés par l'utilisateur et présent dans la base de données XLSTAT.

Onglet **Sorties** :

Bilan de l'optimisation : activez cette option pour afficher le bilan de l'optimisation effectuée pour générer le plan d'expériences.

Tableau de Burt : activez cette option pour afficher le tableau de Burt associé au plan d'expériences.

Plan codé : activez cette option pour afficher le tableau du plan d'expériences encodé dans le cas d'un plan d-optimal.

Essai en plein champs : activez cette option pour afficher le plan sous forme de parcelles.

Afficher les feuilles d'expérience : activez cette option si vous désirez afficher des feuilles Excel individuelles pour chaque expérience. Ceci peut être utile afin de les imprimer et de réaliser les expériences.

Trier en ordre croissant : activez cette option pour trier les modalités comparées en ordre croissant, le critère de tri étant leur moyenne respective. Si cette option n'est pas activée, le tri est décroissant.

Trier alphabétiquement : activez cette option si vous voulez que XLSTAT trie alphabétiquement les noms des modalités.

Libellés Variable-Modalité : activez cette option pour que les libellés des lignes et des colonnes du tableau de contingence utilisent le nom de la variable suivi du nom des modalités. Si cette option n'est pas activée, les libellés sont construits uniquement à partir des noms des modalités.

Onglet **Graphiques** :

Vue 3D du tableau de Burt : activez cette option pour afficher le graphique 3D du tableau de Burt.

Boîte de dialogue **Plans d'effet de facteurs / Plans communs** :

Sélection du plan d'expériences : cette boîte de dialogue permet à l'utilisateur de sélectionner le plan d'expériences à utiliser. Une liste de plans factoriels fractionnaires est affichée avec la distance au plan demandé. Si vous sélectionnez l'un des plans et cliquez sur le bouton sélectionner, alors le plan sélectionné sera affiché. Si aucun des plans ne vous convient, cliquez sur le bouton optimiser et l'algorithme de recherche d'un plan D-optimal est lancé.

Résultats

Information sur les variables : ce tableau récapitule les informations sur les facteurs. Pour chaque facteur, le nom court, le nom long et l'unité utilisée sont affichés.

Plan d'expériences : le plan d'expériences est affiché dans ce tableau. Des colonnes supplémentaires contenant des informations sur les facteurs et sur les réponses, ainsi qu'un libellé pour chaque expérience, l'ordre de tri, l'ordre d'exécution et le numéro de la répétition sont aussi inclus dans ce tableau.

Optimisation des réponses : le tableau d'optimisation des réponses est affiché à la suite du plan d'expériences. Vous devez alors sélectionner les paramètres suivants :

- **Objectif** : choisissez l'objectif de l'optimisation. Vous avez le choix entre minimum, optimum et maximum.

Si l'objectif sélectionné est l'optimum ou le maximum, les champs suivants sont activés :

- **Inférieur** : entrez pour chaque réponse la valeur de la borne inférieure en dessous de laquelle la désirabilité vaut 0.
- **Cible (gauche)** : entrez pour chaque réponse la valeur de la borne inférieure au dessus de laquelle la désirabilité vaut 1.

Si l'objectif sélectionné est l'optimum ou le minimum, les champs suivants sont activés :

- **Cible (droite)** : entrez pour chaque réponse la valeur de la borne supérieure en dessous de laquelle la désirabilité vaut 1.

- **Inférieur** : entrez pour chaque réponse la valeur de la borne supérieure en dessus de laquelle la désirabilité vaut 0.

- **s** : activez cette option si la fonction de désirabilité croissante doit être non linéaire. Entrez alors la valeur du paramètre de forme qui doit se trouver entre 0,01 et 100.

- **t** : activez cette option si la fonction de désirabilité décroissante doit être non linéaire. Entrez alors la valeur du paramètre de forme qui doit se trouver entre 0,01 et 100.

- **Poids** : activez cette option si les réponses doivent avoir une valeur exponentielle différente de 1 lors du calcul de la désirabilité. Entrez alors la valeur du paramètre de forme qui doit se trouver entre 0,01 et 100.

Pour plus de détails sur l'optimisation des réponses, vous pouvez vous référez à l'aide de l'analyse d'un plan d'effet de facteurs

Plan codé : le plan encodé est affiché. Ce tableau n'est disponible que si le plan d'expériences est un plan d-optimal.

Tableau de Burt : le tableau de Burt est affiché si l'option correspondante a été activée dans la boîte de dialogue. Une visualisation 3D de ce tableau est aussi affichée si l'option est activée dans l'onglet « Graphiques » de la boîte de dialogue.

Si l'option d'optimisation a été active, les résultats suivants sont affichés:

Bilan de l'optimisation : un tableau donnant le nombre d'expérience, les critères du $log(déterminant)$ le critère $norm.log(déterminant)$ et le critère $Log(|I|^{1/p})$ est affiché. Le meilleur résultat est en gras.Ensuite le critère $norm.log(déterminant)$ est affiché dans un diagramme en fonction du nombre d'expériences.

Statistiques pour chaque itération : ce tableau affiche, pour le plan sélectionné, l'évolution du critère d'optimisation durant les itérations effectuées. Un graphique montrant cette évolution est aussi disponible si l'option correspondante est activée dans l'onglet « Graphiques ».

De plus, un tableau donnant le nombre d'expérience, le nombre d'itérations pour l'optimisation, les critères du log(déterminant) et le critère norm.log(déterminant) est affiché. Le meilleur résultat est en gras.

Si l'option de génération de feuilles d'expériences a été activée et s'il y a moins de 200 expériences, une feuille Excel pour chaque expérience est ajoutée dans le classeur Excel.

Ces feuilles individuelles débutent par le titre et le nom du modèle afin de simplifier leur identification. Ensuite, le numéro de l'expérience ainsi que le nombre total d'expériences sont affichés. Les informations présentes dans les colonnes supplémentaires du tableau d'expérience sont aussi affichées.

Finalement, les informations sur les valeurs des facteurs sont affichées de manière à ce que l'utilisateur puisse entrer les résultats pour les différentes réponses.

Ces feuilles peuvent être imprimées ou utilisées sous forme de feuille Excel durant la réalisation des expériences.

Exemple

Un exemple de génération et d'analyse d'un plan d'effet de facteurs est disponible sur le Centre d'aide XLSTAT à l'adresse suivante :

http://www.xlstat.com/demo-doe1f.htm

Bibliographie

Louvet, F. and Delplanque L. (2005). Design Of Experiments: The French touch, Les plans d'expériences : une approche pragmatique et illustrée, Alpha Graphic, Olivet, 2005.

Montgomery D.C. (2005), Design and Analysis of Experiments, 6th édition, John Wiley & Sons.

Myers, R. H., Khuri, I. K. and Carter W. H. Jr. (1989). Response Surface Methodology: 1966 – 1988, *Technometrics*, **31**, 137-157.

Analyse d'un plan d'effet de facteurs

Utilisez cet outil pour analyser un plan d'effet de facteurs. Un modèle du second ordre est utilisé pour cette analyse.

Dans cette section :

Description

L'analyse d'un plan d'effet de facteurs utilise les mêmes méthodes que dans une régression linéaire ou une analyse de la variance (ANOVA), c'est-à-dire le modèle linéaire général. L'unique différence vient du type de variables explicatives utilisées. Dans une analyse de plan d'effet de facteurs, on utilise des variables qualitatives ou facteurs comme dans le cas de l'ANOVA.

Si p est le nombre de facteurs, le modèle de l'ANOVA s'écrit de la manière suivante :

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_{k(i,j),j} + \epsilon_i \qquad (1)$$

où Y_i est la valeur observée pour la variable dépendante pour l'observation i, $k(i,j)$ est l'indice correspondant à la modalité du facteur j pour l'observation i, et ϵ_i est l'erreur du modèle.

Les hypothèses utilisées en ANOVA sont identiques à celles de la régression linéaire : les erreurs ϵ_i suivent une même loi normale $\mathcal{N}(0, s)$ et sont indépendantes.

L'écriture du modèle complétée par cette hypothèse a pour conséquence que, dans le cadre du modèle de régression linéaire, les Y_i sont des réalisations de variables aléatoires de moyenne μ_i et de variance s^2, avec

$$\mu_i = \beta_0 + \sum_{j=1}^{p} \beta_{k(i,j),j}$$

Si l'on souhaite utiliser les différents tests proposés dans les résultats de la régression linéaire il est recommandé de vérifier a posteriori que les hypothèses sous-jacentes sont bien vérifiées. La normalité des résidus peut être vérifiée en analysant certains graphiques ou en utilisant un

test de normalité. L'indépendance des résidus peut être vérifiée en analysant certains graphiques ou en utilisant le test de Durbin Watson.

Pour plus de détails sur l'ANOVA et la régression linéaire, veuillez vous référer aux chapitres de l'aide à ce sujet.

Optimisation des réponses et désirabilité

Il est possible d'optimiser chaque réponse de manière individuelle et de combiner les résultats afin d'obtenir une fonction de désirabilité et d'analyser ses valeurs. Introduite par Derringer and Suich (1980), cette approche est basée sur la transformation de la réponse y_i en une fonction de désirabilité individuelle d_i sur l'intervalle $0 \leq d_i \leq 1$.

Lorsque y_i a atteint sa valeur cible, alors $d_i = 1$. Si y_i se trouve en dehors d'une région acceptable autour de la cible, alors $d_i = 0$. Entre ces deux cas, des valeurs intermédiaires de d_i sont obtenues.

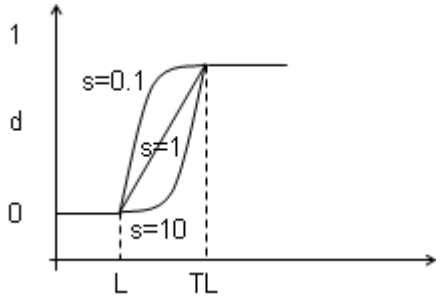
Les 3 différents cas pour d_i sont :

- L = borne inférieure. Pour chaque valeur $< L$, on a $d_i = 0$
- U = borne supérieure. Pour chaque valeur $> U$, on a $d_i = 0$.
- $T(L)$ = valeur cible à gauche.
- $T(R)$ = valeur cible à droite. Pour chaque valeur entre $T(L)$ et $T(R)$, on a $d_i = 1$.
- s, t = paramètres permettant de définir la forme de la fonction d'optimisation entre L et $T(L)$ et entre $T(R)$ et U.

L'équation suivante doit être respectée lorsqu'on définit L, U, $T(L)$ et $T(R)$:

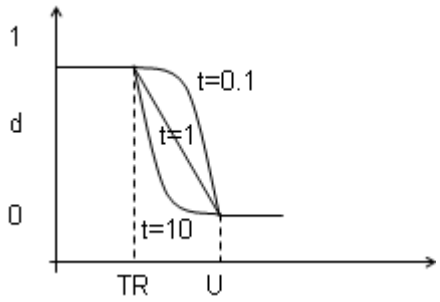
$$L \leq T(L) \leq T(R) \leq U$$

Maximiser la valeur de y_i:



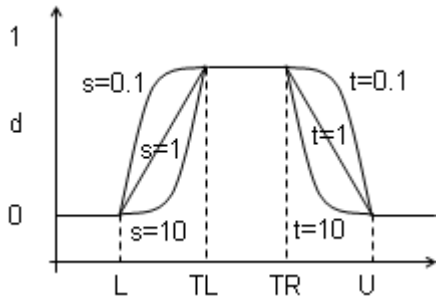
$$d_i = \begin{cases} 0 & y_i < L \\ (\frac{y_i - L}{T(L) - L})^s & L \leq y_i \leq T(L) \\ 1 & y_i > T(L) \end{cases}$$

Minimiser la valeur de y_i:



$$d_i = \begin{cases} 1 & y_i < T(R) \\ (\frac{U-y_i}{U-T(R)})^t & T(R) \leq y_i \leq U \\ 0 & y_i > U \end{cases}$$

Afin de cibler un intervalle donné de y_i, on peut utiliser la fonction de désirabilité suivante :



$$d_i = \begin{cases} 0 & y_i < L \\ (\frac{y_i-L}{T(L)-L})^s & L \leq y_i \leq T(L) \\ 1 & T(L) < y_i < T(R) \\ (\frac{U-y_i}{U-T(R)})^t & T(R) \leq y_i \leq U \\ 0 & y_i > U \end{cases}$$

Les paramètres sont choisis de manière à maximiser la désirabilité globale D.

$$D = (d_1^{w_1} \cdot d_2^{w_2} \cdot ... \cdot d_m^{w_m})^{\frac{1}{w_1 \cdot w_2 \cdot ... \cdot w_m}}$$

Où $1 \leq w_i \leq 10$ sont des poids associés aux fonctions de désirabilité individuelles. Plus les w_i sont grands, plus les d_i sont pris en compte lors de l'optimisation.

Au moment de l'affichage, XLSTAT donne les 5 meilleures solutions trouvées lors de l'optimisation.

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.

OK : cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.


: cliquez sur ce bouton pour changer la façon dont vous chargez les données dans XLSTAT. Si le bouton possède une icône de souris, XLSTAT vous permet de faire une sélection à la souris des données. Si le bouton possède une icône de liste, XLSTAT vous permet de sélectionner les données à partir d'une liste.

Onglet **Général** :

Y / résultats : sélectionnez les colonnes du plan d'expériences correspondant aux réponses. Ces colonnes doivent maintenant être remplies avec les résultats des expériences qui ont été menées.

Plan d'expériences : sélectionnez votre plan d'expériences. Si vous avez modifié votre plan, vérifiez que les facteurs qualitatifs et quantitatifs se suivent. L'ensemble des colonnes du plan doivent être sélectionnées.

Nombre de facteurs quantitatifs : entrez le nombre de facteurs quantitatifs de votre plan d'expériences.

Nombre de facteurs qualitatifs : entrez le nombre de facteurs qualitatifs de votre plan d'expériences.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des variables : activez cette option si la première ligne des données sélectionnées contient un libellé.

Onglet **Réponses** :

Optimisation des réponses : activez cette option si vous souhaitez réaliser une optimisation des réponses. Dans ce cas sélectionnez le tableau d'optimisation des réponses généré au moment de la création du plan. L'en-tête du tableau doit être inclus dans la sélection.

- **Objectif** : choisissez l'objectif de l'optimisation. Vous avez le choix entre minimum, optimum et maximum.

Si l'objectif sélectionné est l'optimum ou le maximum, les champs suivants sont activés :

- **Inférieur** : entrez pour chaque réponse la valeur de la borne inférieure en dessous de laquelle la désirabilité vaut 0.
- **Cible (gauche)** : entrez pour chaque réponse la valeur de la borne inférieure au dessus de laquelle la désirabilité vaut 1.

Si l'objectif sélectionné est l'optimum ou le minimum, les champs suivants sont activés :

- **Cible (droite)** : entrez pour chaque réponse la valeur de la borne supérieure en dessous de laquelle la désirabilité vaut 1.
- **Inférieur** : entrez pour chaque réponse la valeur de la borne supérieure en dessus de laquelle la désirabilité vaut 0.
- **s** : activez cette option si la fonction de désirabilité croissante doit être non linéaire. Entrez alors la valeur du paramètre de forme qui doit se trouver entre 0,01 et 100.
- **t** : activez cette option si la fonction de désirabilité décroissante doit être non linéaire. Entrez alors la valeur du paramètre de forme qui doit se trouver entre 0,01 et 100.
- **Poids** : activez cette option si les réponses doivent avoir une valeur exponentielle différente de 1 lors du calcul de la désirabilité. Entrez alors la valeur du paramètre de forme qui doit se trouver entre 0,01 et 100.

Interactions / Niveau : activez cette option pour inclure des interactions dans le modèle puis entrez le niveau maximum d'interaction (valeur comprise entre 1 et 4).

Onglet **Sorties** :

Analyse de la variance : activez cette option pour afficher le tableau d'analyse de la variance.

Coefficients d'ajustement : activez cette option pour afficher le tableau des indices de qualité d'ajustement du modèle.

Contributions : activez cette option pour afficher le tableau des contributions. Cette option est nécessaire si vous souhaitez afficher le diagramme de Pareto.

Coefficients normalisés : activez cette option pour afficher les paramètres normalisés du modèle (coefficients bêta).

Prédictions et résidus : activez cette option pour afficher les prédictions et les résidus pour l'ensemble des observations.

- **Prédictions ajustées** : activez cette option pour calculer et afficher les prédictions ajustées dans le tableau des prédictions et résidus.

- **D de Cook** : activez cette option pour calculer et afficher les distances de Cook dans le tableau des prédictions et résidus.
- **Résidus studentisés** : activez cette option pour calculer et afficher les résidus studentisés dans le tableau des prédictions et résidus.

Moyennes : activez cette option pour calculer et afficher les moyennes des modalités des variables qualitatives.

- **LS means** : activez cette option pour utiliser les LS means (Least square means) estimées à partir du modèle et non des observations.
- **Erreurs standard** : activez cette option pour calculer et afficher les erreurs standard associées aux moyennes.
- **Intervalles de confiance** : activez cette option pour calculer les intervalles de confiance autour des moyennes.

Onglet **Graphiques** :

Graphiques de régression : activez cette option pour afficher les graphiques de régression :

- **Coefficients normalisés** : activez cette option pour afficher sur un graphique les paramètres normalisés du modèle avec leur intervalle de confiance.
- **Prédictions et résidus** : activez cette option pour afficher les graphiques suivants :

(1) Droite de régression : ce graphique n'est affiché que s'il n'y a qu'une seule variable explicative, et que cette variable est quantitative.

(2) Variable explicative versus résidus normalisés : ce graphique n'est affiché que s'il n'y a qu'une seule variable explicative, et que cette variable est quantitative.

(3) Variable dépendante versus résidus normalisés.

(4) Prédictions pour la variable dépendante versus variable dépendante.

(5) Graphique en bâtons des résidus normalisés.

Intervalles de confiance : activez cette option pour afficher les intervalles de confiance sur les graphiques (1) et (4).

Diagrammes de Pareto : activez cette option pour afficher le diagramme de Pareto pour l'ensemble des facteurs.

Graphiques des moyennes : activez cette option pour afficher les graphiques permettant de visualiser les moyennes pour les différentes modalités des différents facteurs.

Résultats

Information sur les variables : ce tableau récapitule les informations sur les facteurs. Pour chaque facteur, le nom court, le nom long et l'unité utilisée sont affichés.

Optimisation des réponses : ce tableau donne les 5 meilleures solutions obtenues lors de l'optimisation des réponses.

Coefficients d'ajustement : dans ce tableau sont affichées les statistiques relatives à l'ajustement du modèle de régression :

- **Observations** : le nombre d'observations prises en compte dans les calculs. Dans les formules présentées ci-dessous n désigne le nombre d'observations.
- **Somme des poids** : la somme des poids des observations prises en compte dans les calculs. Dans les formules présentées ci-dessous W désigne la somme des poids.
- **DDL** : le nombre de degrés de liberté pour le modèle retenu (correspondant à la partie erreurs).
- **R²** : le coefficient de détermination du modèle. Ce coefficient, dont la valeur est comprise entre 0 et 1, n'est affiché que si la constante du modèle n'est pas fixée par l'utilisateur. Sa valeur est définie par

$$R^2 = 1 - \frac{\sum_{i=1}^{n} w_i (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} w_i (y_i - \bar{y})^2} \text{ avec } \bar{y} = \frac{1}{n} \sum_{i=1}^{n} w_i y_i$$

- Le R^2 s'interprète comme la proportion de la variabilité de la variable dépendante expliquée par le modèle. Plus le R^2 est proche de 1, meilleur est le modèle. L'inconvénient du R^2 est qu'il ne prend pas en compte le nombre de variables utilisées pour ajuster le modèle.
- **R²ajusté** : le coefficient de détermination ajusté du modèle. Le R^2 ajusté peut être négatif si le R^2 est voisin de zéro. Ce coefficient n'est affiché que si la constante du modèle n'est pas fixée par l'utilisateur. Sa valeur est définie par

$$\hat{R}^2 = 1 - (1 - R^2) \frac{W - 1}{W - p - 1}$$

Le R^2 ajusté est une correction du R^2 qui permet de prendre en compte le nombre de variables utilisées dans le modèle.

- **MCE** : la moyenne des carrés des erreurs (MCE) est définie par :

$$MCE = \frac{1}{W - p^*} \sum_{i=1}^{n} w_i (y_i - \hat{y}_i)^2$$

- **RMCE** : la racine de la moyenne des carrés des erreurs (RMCE) est la racine carrée de la MCE.
- **MAPE** : la *Mean Absolute Percentage Error* est calculée comme suit :

$$MAPE = \frac{1}{W - p^*} \sum_{i=1}^{n} w_i \left| \frac{y_i - \hat{y}_i}{y_i} \right|$$

- **DW** : le coefficient de Durbin-Watson est défini par

$$DW = \frac{\sum_{i=2}^{n} [(y_i - \hat{y}_i) - (y_{i-1} - \hat{y}_{i-1})]^2}{\sum_{i=1}^{n} w_i (y_i - \hat{y}_i)^2}$$

Ce coefficient correspond au coefficient d'autocorrélation d'ordre 1 et permet de vérifier que les résidus du modèle ne sont pas autocorrélés, sachant que l'indépendance des résidus est l'une des hypothèses de base de la régression linéaire. L'utilisateur pourra se référer à une table des coefficients de Durbin-Watson pour vérifier si l'hypothèse d'indépendance des résidus est acceptable.

- **Cp** : le coefficient Cp de Mallows est défini par

$$Cp = \frac{SCE}{\hat{\sigma}} + 2p^* - W$$

où SCE est la somme du carré des erreurs pour le modèle avec p variables explicatives, et où $\hat{\sigma}$ correspond à l'estimateur de la variance des résidus pour le modèle comprenant toutes les variables explicatives. Plus le coefficient Cp est proche de p^* moins le modèle est biaisé.

- **AIC** : le critère d'information d'Akaike (Akaike's Information Criterion) est défini par

$$AIC = W \ln(\frac{SCE}{W}) + 2p^*$$

Ce critère proposé par Akaike (1973) dérive de la théorie de l'information, et s'appuie sur la mesure de Kullback et Leibler (1951). C'est un critère de sélection de modèles qui pénalise les modèles pour lesquels l'ajout de nouvelles variables explicatives n'apporte pas suffisamment d'information au modèle, l'information étant mesurée au travers de la SCE. On cherche à minimiser le critère AIC.

- **SBC** : le critère bayésien de Schwarz (Schwarz's Bayesian Criterion) est défini par

$$SBC = W \ln(\frac{SCE}{W}) + \ln(W)p^*$$

Ce critère proposé par Schwarz (1978) est proche du critère AIC, et comme ce dernier on cherche à le minimiser.

- **PC** : le critère de prédiction d'Amemiya (Amemiya's Prediction Criterion) est défini par

$$PC = \frac{(1 - R^2)(W + p^*)}{W - p^*}$$

Ce critère proposé par Amemiya (1980) permet comme le R^2 ajusté de tenir compte de la parcimonie du modèle.

- **Press** : la statistique du Press (predicted residual error sum of squares) n'est affichée que si l'option correspondante a été activée dans la boîte de dialogue. Elle est définie par

$$Press = \sum_{i=1}^{n} w_i (y_i - \hat{y}_{i(-i)})^2$$

où $\hat{y}_{i(-i)}$ est la prédiction pour l'observation i lorsque cette dernière n'est pas utilisée pour l'estimation des paramètres. On obtient alors

$$\text{Press RMCE } = \sqrt{\frac{\text{Press}}{W - p^*}}$$

Le Press RMCE peut alors être comparé au RMCE. Une différence importante entre les deux indique que le modèle est sensible à la présence ou absence de certaines observations dans le modèle.

- **Q²** : La statistique Q^2 est affichée. Elle est définie par :

$$Q^2 = 1 - \frac{PressRMSE}{SSE}$$

Plus cet indice est proche de 1, plus le modèle et bon et robuste.

Le **tableau d'analyse de la variance** permet d'évaluer le pouvoir explicatif des variables explicatives. Dans le cas où la constante du modèle n'est pas fixée à une valeur donnée, le pouvoir explicatif est évalué en comparant l'ajustement (au sens des moindres carrés) du modèle final avec l'ajustement du modèle rudimentaire composé d'une constante égale à la moyenne de la variable dépendante. Dans le cas où la constante du modèle est fixée, la comparaison est faite par rapport au modèle pour lequel la variable dépendante serait égale à la constante fixée.

Si l'option Type I/III SS (SS : Sum of Squares) est activée, les tableaux suivants sont affichés.

Le tableau des **Type I SS** permet de visualiser l'influence de l'ajout progressif des variables explicatives sur l'ajustement du modèle, au sens de la somme des carrés des erreurs (SCE), de la moyenne des carrés des erreurs (MCE), du F de Fisher, ou de la probabilité associée au F de Fisher. Plus la probabilité est faible, plus la contribution de la variable au modèle est importante, toutes les autres variables étant déjà dans le modèle. Remarques : l'ordre de sélection des variables dans le modèle influe sur les valeurs obtenues ; la somme des sommes des carrés de ce tableau est égal à la somme des carrés du modèle.

Le tableau des **Type II SS** permet de visualiser l'influence du retrait d'une variable explicative sur l'ajustement du modèle, toutes les autres variables étant conservées, au sens de la somme des carrés des erreurs (SCE), de la moyenne des carrés des erreurs (MCE), du F de Fisher, ou de la probabilité associée au F de Fisher. Plus la probabilité est faible, plus la contribution de la variable au modèle est importante. Remarque : dans le cas des ANOVAs déséquilibrées, l'utilisation des Type III est recommandée mais XLSTAT affiche les Type II pour les utilisateurs avancés qui voudraient disposer des Type II.

Le tableau des **Type III SS** permet de visualiser l'influence du retrait d'une variable explicative sur l'ajustement du modèle, toutes les autres variables étant conservées, au sens de la somme des carrés des erreurs (SCE), de la moyenne des carrés des erreurs (MCE), du F de Fisher, ou de la probabilité associée au F de Fisher. Plus la probabilité est faible, plus la contribution de la variable au modèle est importante, toutes les autres variables étant déjà dans le modèle. Remarque : contrairement au cas des Type I SS, l'ordre de sélection des variables dans le modèle n'influe pas sur les valeurs obtenues, et contrairement aux Type II SS, les valeurs ne dépendent pas des effectifs des cellules (par cellule on entend une combinaison de modalités des différents facteurs), ce qui fait des Type III le test recommandé pour évaluer la contribution d'une variable.

Le tableau **paramètres du modèle** affiche l'estimation des paramètres, l'erreur type correspondante, le t de Student, la probabilité correspondante, ainsi que l'intervalle de confiance

L'**équation du modèle** est ensuite affichée pour faciliter la lecture ou la réutilisation du modèle.

Le tableau des **coefficients normalisés** (aussi appelés coefficients bêta) permet de comparer le poids relatif des variables. Plus la valeur absolue d'un coefficient est élevée, plus le poids de la variable correspondante est important. Lorsque l'intervalle de confiance autour des coefficients normalisés comprend la valeur 0 (cela est facilement visible sur le graphique des coefficients normalisés), le poids d'une variable dans le modèle n'est pas significatif.

Dans le tableau des **prédictions et résidus** sont donnés pour chaque observation, son poids, la valeur de la variable explicative qualitative s'il n'y en a qu'une, la valeur observée de la variable dépendante, la prédiction du modèle, les résidus, les intervalles de confiance, ainsi que la prédiction ajustée et le D de Cook si les options correspondantes ont été activées dans la boîte de dialogue. Deux types d'intervalles de confiance sont affichés : un intervalle de confiance autour de la moyenne (correspondant au cas où l'on ferait la prédiction pour un nombre infini d'observations avec un ensemble de valeurs données des variables explicatives) et un intervalle autour de la prédiction ponctuelle (correspondant au cas d'une prédiction isolée pour des valeurs données des variables explicatives). Le second intervalle est toujours plus grand que le premier, les aléas étant plus importants. Si des données de validation ont été sélectionnées, elles sont affichées en fin de tableau.

Les **graphiques** qui suivent permettent de visualiser les résultats mentionnés ci-dessus. S'il n'y a qu'une seule variable explicative dans le modèle, le premier graphique affiché permet de visualiser les valeurs observées, la droite de régression et les deux types d'intervalles de confiance autour des prévisions. Le second graphique permet quant à lui de visualiser les résidus normalisés en fonction de la variable explicative. En principe, les résidus doivent être distribués de manière aléatoire autour de l'axe des abscisses. L'observation d'une tendance ou d'une forme révèlerait un problème au niveau du modèle.

Les **trois graphiques** affichés ensuite permettent de visualiser respectivement l'évolution des résidus normalisés en fonction de la variable dépendante, la distance entre les prédictions et les observations (pour un modèle idéal, les points seraient tous sur la bissectrice), et les résidus normalisés sur la forme d'un diagramme en bâtons. Ce dernier graphique permet de rapidement voir si un nombre anormal de données sort de l'intervalle $]-2,2[$ sachant que ce dernier, sous hypothèse de normalité, doit contenir environ 95% des données.

Exemple

Un exemple de génération et d'analyse d'un plan de surface de réponse est disponible sur le Centre d'aide XLSTAT à l'adresse suivante :

http://www.xlstat.com/demo-doe3f.htm

Bibliographie

Derringer R. and Suich R. (1980). Simultaneous optimization of several response variables, *Journal of Quality Technoloty*, **12**, 214-219.

Louvet, F. and Delplanque L. (2005). Design Of Experiments: The French touch, Les plans d'expériences : une approche pragmatique et illustrée, Alpha Graphic, Olivet, 2005.

Montgomery D.C. (2005), Design and Analysis of Experiments, 6th édition, John Wiley & Sons.

Myers, R. H., Khuri, I. K. and Carter W. H. Jr. (1989). Response Surface Methodology: 1966 – 1988, *Technometrics*, **31**, 137-157.

Plans de surface de réponse

Utilisez ce module pour générer des plans de surface de réponse pour 2 à 10 facteurs et au moins une réponse.

Dans cette section :

Description

La famille des plans de surface de réponse est utilisée pour modéliser et analyser des problèmes dans lesquels la réponse est influencée par plusieurs variables et pour lesquels l'objectif est d'optimiser la réponse.

Remarque : Les plans d'effet de facteurs visent, quant à eux, à étudier les facteurs et non la réponse.

Par exemple, dans un processus industriel, supposons qu'un ingénieur cherche la pression optimale (x_1) et la température optimale (x_2) qui lui permette d'obtenir une solidité maximale y. On peut alors écrire l'équation suivante :

$$y = f(x_1, x_2) + \epsilon_i \quad (1)$$

L'ingénieur utilisera un plan de surface de réponse afin de générer ses expériences qui lui permettront de trouver les valeurs optimales de x_1 et x_2.

Modèle

Cet outil est basé sur un modèle du second ordre. Si k est le nombre de facteurs, le modèle quadratique s'écrit de la manière suivante :

$$Y = \beta_0 + \sum_{i=1}^{k} \beta_i x_i + \sum_{i=1}^{k} \beta_{ii} x_i^2 + \sum \sum \beta_{ij} x_i x_j + \epsilon \quad (2)$$

Plan d'expérience

Cet outil propose les plans suivants pour la modélisation de surfaces :

Matrice grille à 2 niveaux (plan factoriel complet à 2 niveaux) : Toutes les combinaisons possibles de 2 valeurs pour chacun des facteurs (le minimum et le maximum) sont générées dans le plan d'expérience. Le nombre d'expériences n pour k facteurs est donné par :

$$n = 2^k$$

Matrice grille à 3 niveaux (plan factoriel complet à 3 niveaux) : Toutes les combinaisons possibles de 3 valeurs pour chacun des facteurs (le minimum, le maximum et la moyenne) sont générées dans le plan d'expérience. Le nombre d'expériences n pour k facteurs est donné par :

$$n = 3^k$$

Plan composite centré : Ce plan d'expérience a été introduit par Box G.E.P. et Wilson K.B. (1951), les points de l'expérience sont générés sur une sphère centrée autour d'un point central. Le nombre de niveaux des différents facteurs est minimisé. Le calcul du point central est répété à plusieurs reprises de façon à maximiser la précision de la prédiction autour du point supposé optimal. Le nombre de répétitions dans la recherche du point central n_0 est calculé en utilisant la formule suivante pour k facteurs :

$$\gamma = \frac{(k+3) + \sqrt{9k^2 + 14k - 7}}{4(k+2)}$$

et

$$n_0 = floor(\gamma(\sqrt{2}+2)^2 - 2^k - 2k)$$

Avec $floor$ qui désigne le plus grand entier plus petit que l'argument entre parenthèses. Le nombre d'expériences pour k facteurs est donné par :

$$n = 2^k + 2k + 1$$

Box-Behnken : Ce plan d'expérience a été introduit par Box G.E.P. et Behnken D.W (1960). Il est basé sur le même principe que le plan composite centré mais nécessite moins d'expériences. Le nombre d'expériences pour k facteurs est donné par :

$$n = 2k^2 - 2k + 1$$

Doehlert : Ce plan d'expérience a été introduit par Doehlert D.H. (1970). Il est basé sur le même principe que le plan composite centré et le plan de Box-Behnken, mais nécessite moins d'expériences. Ce plan utilise plus de niveaux de facteurs et pourra être plus complexe à traiter dans la pratique. Le nombre d'expériences pour k facteurs est donné par :

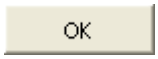
$$n = k^2 + k + 1$$

Le tableau suivant montre le nombre d'expériences pour chacun des 4 types de plans en fonction du nombre k de facteurs. Dans les calculs, le point central n'est présent qu'une seule fois.

k	full fact.	Cent. comp.	Box-Behnken	Doehlert
2	9	9	5	7
3	27	15	13	13
4	81	25	25	21
5	243	43	41	31
6	729	77	61	43
7	2187	143	85	57

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.


: cliquez sur ce bouton pour changer la façon dont vous chargez les données dans XLSTAT. Si le bouton possède une icône de souris, XLSTAT vous permet de faire une sélection à la souris des données. Si le bouton possède une icône de liste, XLSTAT vous permet de sélectionner les données à partir d'une liste.

Onglet **Général** :

Facteurs quantitatifs (min/max) : sélectionnez le minimum et le maximum des facteurs quantitatifs. La sélection doit comporter deux lignes, la première correspond aux minimums, la seconde aux maximums pour chacun des facteurs.

Facteurs qualitatifs : sélectionnez le tableau des facteurs qualitatifs et leurs modalités.

Nombre de réponses : entrez le nombre de réponses à analyser.

Répétitions : activez cette option afin de sélectionner le nombre de répétitions du plan d'expériences.

Ordre aléatoire : activez cette option si vous souhaitez que l'ordre des expériences dans le plan soit aléatoire.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des variables : activez cette option si la première ligne de la sélection contient le libellé des variables.

Onglet **Options** :

Plan d'expérience : choisissez le plan d'expérience que vous désirez utiliser. En fonction du nombre de facteurs, différents plans sont proposés.

Nombre de points centraux : dans le cas d'un plan composite centré, vous avez la possibilité de changer le nombre de répétitions du point central. Activez cette option pour entrer une valeur pour ce nombre.

Onglet **Sorties** :

Plan codé : activez cette option pour afficher le tableau du plan d'expériences encodé.

Afficher les feuilles d'expérience : activez cette option si vous désirez afficher des feuilles Excel individuelles pour chaque expérience. Ceci peut être utile afin de les imprimer et de réaliser les expériences.

Résultats

Information sur les variables : ce tableau regroupe les informations sur les facteurs. Pour chaque facteur, le nom court, le nom long et l'unité utilisée sont affichés.

Plan d'expériences : le plan d'expériences est affiché dans ce tableau. Des colonnes supplémentaires contenant des informations sur les facteurs et sur les réponses, ainsi qu'un libellé pour chaque expérience, l'ordre de tri, l'ordre d'exécution et le numéro de la répétition sont aussi inclus dans ce tableau.

Optimisation des réponses : le tableau d'optimisation des réponses est affiché à la suite du plan d'expériences. Vous devez alors sélectionner les paramètres suivants :

- **Objectif** : choisissez l'objectif de l'optimisation. Vous avez le choix entre minimum, optimum et maximum.

Si l'objectif sélectionné est l'optimum ou le maximum, les champs suivants sont activés :

- **Inférieur** : entrez pour chaque réponse la valeur de la borne inférieure en dessous de laquelle la désirabilité vaut 0.
- **Cible (gauche)** : entrez pour chaque réponse la valeur de la borne inférieure au dessus de laquelle la désirabilité vaut 1.

Si l'objectif sélectionné est l'optimum ou le minimum, les champs suivants sont activés :

- **Cible (droite)** : entrez pour chaque réponse la valeur de la borne supérieure en dessous de laquelle la désirabilité vaut 1.
- **Inférieur** : entrez pour chaque réponse la valeur de la borne supérieure au dessus de laquelle la désirabilité vaut 0.
- **s** : activez cette option si la fonction de désirabilité croissante doit être non linéaire. Entrez alors la valeur du paramètre de forme. Cette valeur doit être comprise entre 0,01 et 100.
- **t** : activez cette option si la fonction de désirabilité décroissante doit être non linéaire. Entrez alors la valeur du paramètre de forme. Cette valeur doit être comprise entre 0,01 et 100.
- **Poids** : activez cette option si les réponses doivent avoir une valeur exponentielle différente de 1 lors du calcul de la désirabilité. Entrez alors la valeur du paramètre de forme. Cette valeur doit être comprise entre 0,01 et 100.

Pour plus de détails sur l'optimisation des réponses, vous pouvez vous référez à l'aide de l'analyse d'un plan d'effet de facteurs

Plan codé : le plan encodé est affiché. Ce tableau n'est disponible que si le plan d'expériences est un plan d-optimal.

Si l'option de génération de feuilles d'expériences a été activée et s'il y a moins de 200 expériences, une feuille Excel pour chaque expérience est ajoutée dans le classeur Excel.

Ces feuilles individuelles débutent par le titre et le nom du modèle afin de simplifier leur identification. Ensuite, le numéro de l'expérience ainsi que le nombre total d'expériences sont affichés. Les informations présentes dans les colonnes supplémentaires du tableau d'expérience sont aussi affichées.

Finalement, les informations sur les valeurs des facteurs sont affichées de manière à ce que l'utilisateur puisse entrer les résultats pour les différentes réponses.

Ces feuilles peuvent être imprimées ou utilisées sous forme de feuille Excel durant la réalisation des expériences.

Exemple

Un exemple de génération et d'analyse d'un plan de surface de réponse est disponible sur le Centre d'aide XLSTAT à l'adresse suivante :

http://www.xlstat.com/demo-doe2f.htm

Bibliographie

Box G. E. P. and Behnken D. W. (1960). Some new three level designs for the study of quantitative variables, *Technometrics*, **2**, Number 4, 455-475.

Box G. E. P. and Wilson K. B. (1951). On the experimental attainment of optimum conditions, *Journal of Royal Statistical Society*, **13**, Serie B, 1-45.

Doehlert D. H. (1970). Uniform shell designs, *Journal of Royal Statistical Society*, **19**, Serie C, 231-239.

Louvet, F. and Delplanque L. (2005). Design Of Experiments: The French touch, Les plans d'expériences : une approche pragmatique et illustrée, Alpha Graphic, Olivet, 2005.

Montgomery D.C. (2005), Design and Analysis of Experiments, 6th édition, John Wiley & Sons.

Myers, R. H., Khuri, I. K. and Carter W. H. Jr. (1989). Response Surface Methodology: 1966 – 1988, *Technometrics*, **31**, 137-157.

Analyse d'un plan de surface de réponse

Utilisez cet outil pour analyser un plan de surface de réponses avec entre 2 et 10 facteurs. Un modèle du second ordre est utilisé pour cette analyse.

Dans cette section :

Description

L'analyse d'un plan de surface de réponse se base sur le même principe qu'une régression linéaire. La différence majeure vient du modèle qui est utilisé pour l'estimation : une fonction quadratique.

Si k est le nombre de facteurs, le modèle quadratique s'écrit de la manière suivante :

$$Y = \beta_0 + \sum_{i=1}^{k} \beta_i x_i + \sum_{i=1}^{k} \beta_{ii} x_i^2 + \sum_{i<j} \sum \beta_{ij} x_i x_j + \epsilon \qquad (1)$$

Pour plus de détails sur l'ANOVA et la régression linéaire, veuillez vous référer aux chapitres de l'aide à ce sujet.

Optimisation des réponses et désirabilité

Il est possible d'optimiser chaque réponse de manière individuelle et de combiner les résultats afin d'obtenir une fonction de désirabilité et d'analyser ses valeurs. Introduite par Derringer and Suich (1980), cette approche est basée sur la transformation de la réponse y_i en une fonction de désirabilité individuelle d_i sur l'intervalle $0 \leq d_i \leq 1$.

Lorsque y_i a atteint sa valeur cible, alors $d_i = 1$. Si y_i se trouve en dehors d'une région acceptable autour de la cible, alors $d_i = 0$. Entre ces deux cas, des valeurs intermédiaires de d_i sont obtenues.

Les 3 différents cas pour d_i sont :

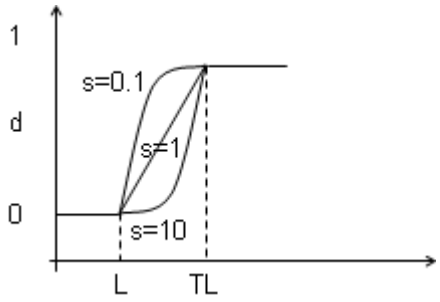
- L = borne inférieure. Pour chaque valeur $< L$, on a $d_i = 0$
- U = borne supérieure. Pour chaque valeur $> U$, on a $d_i = 0$.

- $T(L)$ = valeur cible à gauche.
- $T(R)$ = valeur cible à droite. Pour chaque valeur entre $T(L)$ et $T(R)$, on a $d_i = 1$.
- s, t = paramètres permettant de définir la forme de la fonction d'optimisation entre L et $T(L)$ et entre $T(R)$ et U.

L'équation suivante doit être respectée lorsqu'on définit L, U, $T(L)$ et $T(R)$:

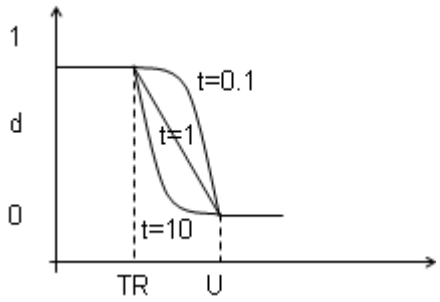
$$L \leq T(L) \leq T(R) \leq U$$

Maximiser la valeur de y_i:



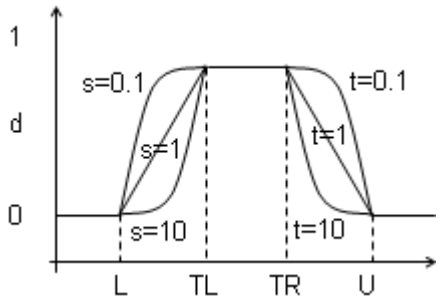
$$d_i = \begin{cases} 0 & y_i < L \\ (\frac{y_i - L}{T(L) - L})^s & L \leq y_i \leq T(L) \\ 1 & y_i > T(L) \end{cases}$$

Minimiser la valeur de y_i:



$$d_i = \begin{cases} 1 & y_i < T(R) \\ (\frac{U - y_i}{U - T(R)})^t & T(R) \leq y_i \leq U \\ 0 & y_i > U \end{cases}$$

Afin de cibler un intervalle donné de y_i, on peut utiliser la fonction de désirabilité suivante :



$$d_i = \begin{cases} 0 & y_i < L \\ (\frac{y_i - L}{T(L) - L})^s & L \leq y_i \leq T(L) \\ 1 & T(L) < y_i < T(R) \\ (\frac{U - y_i}{U - T(R)})^t & T(R) \leq y_i \leq U \\ 0 & y_i > U \end{cases}$$

Les paramètres sont choisis de manière à maximiser la désirabilité globale D.

$$D = (d_1^{w_1} \cdot d_2^{w_2} \cdot \ldots \cdot d_m^{w_m})^{\frac{1}{w_1 \cdot w_2 \cdot \ldots \cdot w_m}}$$

Où $1 \leq w_i \leq 10$ sont des poids associés aux fonctions de désirabilité individuelles. Plus les w_i sont grands, plus les d_i sont pris en compte lors de l'optimisation.

Au moment de l'affichage, XLSTAT donne les 5 meilleures solutions trouvées lors de l'optimisation.

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.

OK : cliquez sur ce bouton pour lancer les calculs.

Annuler : cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

Aide : cliquez sur ce bouton pour afficher l'aide.

: cliquez sur ce bouton pour rétablir les options par défaut.

: cliquez sur ce bouton pour effacer les sélections de données.

: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.

 : cliquez sur ce bouton pour changer la façon dont vous chargez les données dans XLSTAT. Si le bouton possède une icône de souris, XLSTAT vous permet de faire une sélection à la souris des données. Si le bouton possède une icône de liste, XLSTAT vous permet de sélectionner les données à partir d'une liste.

Onglet **Général** :

Y / résultats : sélectionnez les colonnes du plan d'expériences correspondant aux réponses. Ces colonnes doivent maintenant être remplies avec les résultats des expériences qui ont été menées.

Plan d'expériences : sélectionnez votre plan d'expériences. Si vous avez modifié votre plan, vérifiez que les facteurs qualitatifs et quantitatifs se suivent. L'ensemble des colonnes du plan doivent être sélectionnées.

Informations des variables : sélectionnez le tableau contenant les différents facteurs et leurs modalités.

Nombre de facteurs quantitatifs : entrez le nombre de facteurs quantitatifs de votre plan d'expériences.

Nombre de facteurs qualitatifs : entrez le nombre de facteurs qualitatifs de votre plan d'expériences.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des variables : activez cette option si la première ligne des données sélectionnées contient un libellé.

Onglet **Réponses** :

Optimisation des réponses : activez cette option si vous souhaitez réaliser une optimisation des réponses. Dans ce cas sélectionnez le tableau d'optimisation des réponses généré au moment de la création du plan. L'en-tête du tableau doit être inclus dans la sélection.

- **Objectif** : choisissez l'objectif de l'optimisation. Vous avez le choix entre minimum, optimum et maximum.

Si l'objectif sélectionné est l'optimum ou le maximum, les champs suivants sont activés :

- **Inférieur** : entrez pour chaque réponse la valeur de la borne inférieure en dessous de laquelle la désirabilité vaut 0.
- **Cible (gauche)** : entrez pour chaque réponse la valeur de la borne inférieure au dessus de laquelle la désirabilité vaut 1.

Si l'objectif sélectionné est l'optimum ou le minimum, les champs suivants sont activés :

- **Cible (droite)** : entrez pour chaque réponse la valeur de la borne supérieure en dessous de laquelle la désirabilité vaut 1.
- **Inférieur** : entrez pour chaque réponse la valeur de la borne supérieure au dessus de laquelle la désirabilité vaut 0.
- **s** : activez cette option si la fonction de désirabilité croissante doit être non linéaire. Entrez alors la valeur du paramètre de forme. Cette valeur doit être comprise entre 0,01 et 100.
- **t** : activez cette option si la fonction de désirabilité décroissante doit être non linéaire. Entrez alors la valeur du paramètre de forme. Cette valeur doit être comprise entre 0,01 et 100.
- **Poids** : activez cette option si les réponses doivent avoir une valeur exponentielle différente de 1 lors du calcul de la désirabilité. Entrez alors la valeur du paramètre de forme. Cette valeur doit être comprise entre 0,01 et 100.

Interactions / Niveau : activez cette option pour inclure des interactions dans le modèle puis entrez le niveau maximum d'interaction (valeur comprise entre 1 et 4).

Onglet **Sorties** :

Analyse de la variance : activez cette option pour afficher le tableau d'analyse de la variance.

Coefficients d'ajustement : activez cette option pour afficher le tableau des indices de qualité d'ajustement du modèle.

Contributions : activez cette option pour afficher le tableau des contributions. Cette option est nécessaire si vous souhaitez afficher le diagramme de Pareto.

Coefficients normalisés : activez cette option pour afficher les paramètres normalisés du modèle (coefficients bêta).

Prédictions et résidus : activez cette option pour afficher les prédictions et les résidus pour l'ensemble des observations.

- **Prédictions ajustées** : activez cette option pour calculer et afficher les prédictions ajustées dans le tableau des prédictions et résidus.
- **D de Cook** : activez cette option pour calculer et afficher les distances de Cook dans le tableau des prédictions et résidus.
- **Résidus studentisés** : activez cette option pour calculer et afficher les résidus studentisés dans le tableau des prédictions et résidus.

Moyennes : activez cette option pour calculer et afficher les moyennes des modalités des variables qualitatives.

- **LS means** : activez cette option pour utiliser les LS means (Least square means) estimées à partir du modèle et non des observations.

- **Erreurs standard** : activez cette option pour calculer et afficher les erreurs standard associées aux moyennes.
- **Intervalles de confiance** : activez cette option pour calculer les intervalles de confiance autour des moyennes.

Onglet **Graphiques** :

Graphiques de régression : activez cette option pour afficher les graphiques de régression :

- **Coefficients normalisés** : activez cette option pour afficher sur un graphique les paramètres normalisés du modèle avec leur intervalle de confiance.
- **Prédictions et résidus** : activez cette option pour afficher les graphiques suivants :

(1) Droite de régression : ce graphique n'est affiché que s'il n'y a qu'une seule variable explicative, et que cette variable est quantitative.

(2) Variable explicative versus résidus normalisés : ce graphique n'est affiché que s'il n'y a qu'une seule variable explicative, et que cette variable est quantitative.

(3) Variable dépendante versus résidus normalisés.

(4) Prédictions pour la variable dépendante versus variable dépendante.

(5) Graphique en bâtons des résidus normalisés.

Intervalles de confiance : activez cette option pour afficher les intervalles de confiance sur les graphiques (1) et (4).

Graphique en lignes de niveaux : activez cette option pour afficher les graphiques représentant la fonction de désirabilité en lignes de niveaux dans le cas d'un modèle à 2 facteurs.

Graphique de la trace : activez cette option pour afficher les graphiques représentant la trace de la fonction de désirabilité pour chacun des facteurs, avec les autres facteurs fixés à leur valeur moyenne.

Résultats

Information sur les variables : ce tableau récapitule les informations sur les facteurs. Pour chaque facteur, le nom court, le nom long et l'unité utilisée sont affichés.

Optimisation des réponses : ce tableau donne les 5 meilleures solutions obtenues lors de l'optimisation des réponses.

Coefficients d'ajustement : dans ce tableau sont affichées les statistiques relatives à l'ajustement du modèle de régression :

- **Observations** : le nombre d'observations prises en compte dans les calculs. Dans les formules présentées ci-dessous n désigne le nombre d'observations.

- **Somme des poids** : la somme des poids des observations prises en compte dans les calculs. Dans les formules présentées ci-dessous W désigne la somme des poids.

- **DDL** : le nombre de degrés de liberté pour le modèle retenu (correspondant à la partie erreurs).

- **R²** : le coefficient de détermination du modèle. Ce coefficient, dont la valeur est comprise entre 0 et 1, n'est affiché que si la constante du modèle n'est pas fixée par l'utilisateur. Sa valeur est définie par

$$R^2 = 1 - \frac{\sum_{i=1}^{n} w_i (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} w_i (y_i - \bar{y})^2} \text{ avec } \bar{y} = \frac{1}{n} \sum_{i=1}^{n} w_i y_i$$

- Le R^2 s'interprète comme la proportion de la variabilité de la variable dépendante expliquée par le modèle. Plus le R^2 est proche de 1, meilleur est le modèle. L'inconvénient du R^2 est qu'il ne prend pas en compte le nombre de variables utilisées pour ajuster le modèle.

- **R²ajusté** : le coefficient de détermination ajusté du modèle. Le R^2 ajusté peut être négatif si le R^2 est voisin de zéro. Ce coefficient n'est affiché que si la constante du modèle n'est pas fixée par l'utilisateur. Sa valeur est définie par

$$\hat{R}^2 = 1 - (1 - R^2) \frac{W - 1}{W - p - 1}$$

Le R^2 ajusté est une correction du R^2 qui permet de prendre en compte le nombre de variables utilisées dans le modèle.

- **MCE** : la moyenne des carrés des erreurs (MCE) est définie par :

$$MCE = \frac{1}{W - p^*} \sum_{i=1}^{n} w_i (y_i - \hat{y}_i)^2$$

- **RMCE** : la racine de la moyenne des carrés des erreurs (RMCE) est la racine carrée de la MCE.

- **MAPE** : la *Mean Absolute Percentage Error* est calculée comme suit :

$$MAPE = \frac{1}{W - p^*} \sum_{i=1}^{n} w_i \left| \frac{y_i - \hat{y}_i}{y_i} \right|$$

- **DW** : le coefficient de Durbin-Watson est défini par

$$DW = \frac{\sum_{i=2}^{n}[(y_i - \hat{y}_i) - (y_{i-1} - \hat{y}_{i-1})]^2}{\sum_{i=1}^{n} w_i(y_i - \hat{y}_i)^2}$$

Ce coefficient correspond au coefficient d'autocorrélation d'ordre 1 et permet de vérifier que les résidus du modèle ne sont pas autocorrélés, sachant que l'indépendance des résidus est l'une des hypothèses de base de la régression linéaire. L'utilisateur pourra se référer à une table des coefficients de Durbin-Watson pour vérifier si l'hypothèse d'indépendance des résidus est acceptable.

- **Cp** : le coefficient Cp de Mallows est défini par

$$Cp = \frac{SCE}{\hat{\sigma}} + 2p^* - W$$

où SCE est la somme du carré des erreurs pour le modèle avec p variables explicatives, et où $\hat{\sigma}$ correspond à l'estimateur de la variance des résidus pour le modèle comprenant toutes les variables explicatives. Plus le coefficient Cp est proche de p^* moins le modèle est biaisé.

- **AIC** : le critère d'information d'Akaike (Akaike's Information Criterion) est défini par

$$AIC = W \ln(\frac{SCE}{W}) + 2p^*$$

Ce critère proposé par Akaike (1973) dérive de la théorie de l'information, et s'appuie sur la mesure de Kullback et Leibler (1951). C'est un critère de sélection de modèles qui pénalise les modèles pour lesquels l'ajout de nouvelles variables explicatives n'apporte pas suffisamment d'information au modèle, l'information étant mesurée au travers de la SCE. On cherche à minimiser le critère AIC.

- **SBC** : le critère bayésien de Schwarz (Schwarz's Bayesian Criterion) est défini par

$$SBC = W \ln(\frac{SCE}{W}) + \ln(W)p^*$$

Ce critère proposé par Schwarz (1978) est proche du critère AIC, et comme ce dernier on cherche à le minimiser.

- **PC** : le critère de prédiction d'Amemiya (Amemiya's Prediction Criterion) est défini par

$$PC = \frac{(1 - R^2)(W + p^*)}{W - p^*}$$

Ce critère proposé par Amemiya (1980) permet comme le R^2 ajusté de tenir compte de la parcimonie du modèle.

- **Press** : la statistique du Press (predicted residual error sum of squares) n'est affichée que si l'option correspondante a été activée dans la boîte de dialogue. Elle est définie par

$$Press = \sum_{i=1}^{n} w_i (y_i - \hat{y}_{i(-i)})^2$$

où $\hat{y}_{i(-i)}$ est la prédiction pour l'observation i lorsque cette dernière n'est pas utilisée pour l'estimation des paramètres. On obtient alors

$$\text{Press RMCE } = \sqrt{\frac{\text{Press}}{W - p^*}}$$

Le Press RMCE peut alors être comparé au RMCE. Une différence importante entre les deux indique que le modèle est sensible à la présence ou absence de certaines observations dans le modèle.

- **Q²** : La statistique Q^2 est affichée. Elle est définie par :

$$Q^2 = 1 - \frac{PressRMSE}{SSE}$$

Plus cet indice est proche de 1, plus le modèle et bon et robuste.

Le **tableau d'analyse de la variance** permet d'évaluer le pouvoir explicatif des variables explicatives. Dans le cas où la constante du modèle n'est pas fixée à une valeur donnée, le pouvoir explicatif est évalué en comparant l'ajustement (au sens des moindres carrés) du modèle final avec l'ajustement du modèle rudimentaire composé d'une constante égale à la moyenne de la variable dépendante. Dans le cas où la constante du modèle est fixée, la comparaison est faite par rapport au modèle pour lequel la variable dépendante serait égale à la constante fixée.

Si l'option Type I/III SS (SS : Sum of Squares) est activée, les tableaux suivants sont affichés.

Le tableau des **Type I SS** permet de visualiser l'influence de l'ajout progressif des variables explicatives sur l'ajustement du modèle, au sens de la somme des carrés des erreurs (SCE), de la moyenne des carrés des erreurs (MCE), du F de Fisher, ou de la probabilité associée au F de Fisher. Plus la probabilité est faible, plus la contribution de la variable au modèle est importante, toutes les autres variables étant déjà dans le modèle. Remarques : l'ordre de sélection des variables dans le modèle influe sur les valeurs obtenues ; la somme des sommes des carrés de ce tableau est égal à la somme des carrés du modèle.

Le tableau des **Type II SS** permet de visualiser l'influence du retrait d'une variable explicative sur l'ajustement du modèle, toutes les autres variables étant conservées, au sens de la somme des carrés des erreurs (SCE), de la moyenne des carrés des erreurs (MCE), du F de Fisher, ou de la probabilité associée au F de Fisher. Plus la probabilité est faible, plus la contribution de la variable au modèle est importante. Remarque : dans le cas des ANOVAs déséquilibrées, l'utilisation des Type III est recommandée mais XLSTAT affiche les Type II pour les utilisateurs avancés qui voudraient disposer des Type II.

Le tableau des **Type III SS** permet de visualiser l'influence du retrait d'une variable explicative sur l'ajustement du modèle, toutes les autres variables étant conservées, au sens de la somme des carrés des erreurs (SCE), de la moyenne des carrés des erreurs (MCE), du F de Fisher, ou de la probabilité associée au F de Fisher. Plus la probabilité est faible, plus la contribution de la

variable au modèle est importante, toutes les autres variables étant déjà dans le modèle. Remarque : contrairement au cas des Type I SS, l'ordre de sélection des variables dans le modèle n'influe pas sur les valeurs obtenues, et contrairement aux Type II SS, les valeurs ne dépendent pas des effectifs des cellules (par cellule on entend une combinaison de modalités des différents facteurs), ce qui fait des Type III le test recommandé pour évaluer la contribution d'une variable.

Le tableau **paramètres du modèle** affiche l'estimation des paramètres, l'erreur type correspondante, le t de Student, la probabilité correspondante, ainsi que l'intervalle de confiance

L'**équation du modèle** est ensuite affichée pour faciliter la lecture ou la réutilisation du modèle.

Le tableau des **coefficients normalisés** (aussi appelés coefficients bêta) permet de comparer le poids relatif des variables. Plus la valeur absolue d'un coefficient est élevée, plus le poids de la variable correspondante est important. Lorsque l'intervalle de confiance autour des coefficients normalisés comprend la valeur 0 (cela est facilement visible sur le graphique des coefficients normalisés), le poids d'une variable dans le modèle n'est pas significatif.

Dans le tableau des **prédictions et résidus** sont donnés pour chaque observation, son poids, la valeur de la variable explicative qualitative s'il n'y en a qu'une, la valeur observée de la variable dépendante, la prédiction du modèle, les résidus, les intervalles de confiance, ainsi que la prédiction ajustée et le D de Cook si les options correspondantes ont été activées dans la boîte de dialogue. Deux types d'intervalles de confiance sont affichés : un intervalle de confiance autour de la moyenne (correspondant au cas où l'on ferait la prédiction pour un nombre infini d'observations avec un ensemble de valeurs données des variables explicatives) et un intervalle autour de la prédiction ponctuelle (correspondant au cas d'une prédiction isolée pour des valeurs données des variables explicatives). Le second intervalle est toujours plus grand que le premier, les aléas étant plus importants. Si des données de validation ont été sélectionnées, elles sont affichées en fin de tableau.

Les **graphiques** qui suivent permettent de visualiser les résultats mentionnés ci-dessus. S'il n'y a qu'une seule variable explicative dans le modèle, le premier graphique affiché permet de visualiser les valeurs observées, la droite de régression et les deux types d'intervalles de confiance autour des prévisions. Le second graphique permet quant à lui de visualiser les résidus normalisés en fonction de la variable explicative. En principe, les résidus doivent être distribués de manière aléatoire autour de l'axe des abscisses. L'observation d'une tendance ou d'une forme révèlerait un problème au niveau du modèle.

Les **trois graphiques** affichés ensuite permettent de visualiser respectivement l'évolution des résidus normalisés en fonction de la variable dépendante, la distance entre les prédictions et les observations (pour un modèle idéal, les points seraient tous sur la bissectrice), et les résidus normalisés sous la forme d'un diagramme en bâtons. Ce dernier graphique permet de rapidement voir si un nombre anormal de données sort de l'intervalle $]-2,2[$ sachant que ce dernier, sous hypothèse de normalité, doit contenir environ 95% des données.

Ensuite, le **graphique en ligne de niveaux** est affiché, si le plan a deux facteurs et l'option correspondante est activée. L'utilisation de ces graphiques permet d'analyser les dépendances de ces deux facteurs simultanément.

Puis le **graphique de la trace** est affiché, si l'option correspondante est activée. Les graphiques des traces montrent pour chaque facteur, la variable de réponse en fonction du

facteur. Tous les autres facteurs étant fixés à leur valeur moyenne.

Exemple

Un exemple de génération et d'analyse d'un plan de surface de réponse est disponible sur le Centre d'aide XLSTAT à l'adresse suivante :

http://www.xlstat.com/demo-doe2f.htm

Bibliographie

Derringer R. and Suich R. (1980). Simultaneous optimization of several response variables, *Journal of Quality Technoloty*, **12**, 214-219.

Louvet, F. and Delplanque L. (2005). Design Of Experiments: The French touch, Les plans d'expériences : une approche pragmatique et illustrée, Alpha Graphic, Olivet, 2005.

Montgomery D.C. (2005), Design and Analysis of Experiments, 6th édition, John Wiley & Sons.

Myers, R. H., Khuri, I. K. and Carter W. H. Jr. (1989). Response Surface Methodology: 1966 – 1988, *Technometrics*, **31**, 137-157.

Plans de mélange

Utilisez ce module pour générer un plan de mélange pour 2 à 6 facteurs.

Dans cette section :

Description

Les plans de mélange permettent de modéliser les résultats d'expériences lorsque celles-ci portent sur l'optimisation de formulation. Le modèle ainsi obtenu est appelé "loi de mélange".

Les plans de mélange se distinguent des plans factoriels par les caractéristiques suivantes :

- Les facteurs étudiés sont des proportions dont la somme est égale à 1.
- La construction du plan d'expériences est soumise à des contraintes car les facteurs ne peuvent évoluer indépendamment les uns des autres (la somme des proportions valant 1).

Domaine expérimental d'un mélange

Quand on n'impose aucune contrainte aux concentrations des n constituants, le domaine expérimental est un simplexe, c'est-à-dire un polyèdre régulier à n sommets dans un espace de dimension $n-1$. Par exemple, pour un mélange de 3 constituants, le domaine expérimental est un triangle équilatéral ; pour 4 constituants, c'est un tétraèdre régulier.

Les techniques de construction de plans de mélange consistent donc à positionner régulièrement les expériences dans le simplexe afin d'optimiser la précision du modèle. Les constructions les plus classiques sont les plans de Scheffé, de Scheffé centré et les plans augmentés.

Si on introduit des contraintes sur les constituants du modèle en définissant une quantité minimale ou une quantité maximale à ne pas dépasser, alors, selon les cas, le domaine expérimental peut être un simplexe, un simplexe inversé (aussi appelé simplexe B) ou encore un polyèdre convexe quelconque. Dans ce dernier cas, les plans simplexes ne sont plus utilisables.

Attention : si le nombre de constituants est important et que les contraintes sont nombreuses sur les constituants, il est possible que le domaine expérimental n'existe pas.

Les réseaux simplexes de Scheffé sont les plans les plus simples à construire. Ils permettent de construire des modèles de degré quelconque m. Ces matrices sont associées à un modèle canonique comportant un nombre élevé de coefficients (Full Canonical Model).

	Degré du modèle		
Constituants	2	3	4
3	6	10	15
4	10	20	35
5	15	35	70
6	21	56	126
8	36	120	330
10	55	220	715

Afin d'améliorer la séquentialité des expériences, Scheffé a proposé d'ajouter des points au centre du domaine d'expérimentation. Ces plans d'expériences sont connus sous le nom de *Simplex Centroid Designs*.

Ces plans de mélange permettent de construire un modèle polynomial réduit, qui ne comporte que des termes produit des constituants. Le nombre d'expériences augmente donc moins vite que dans le cas d'un simplexe de Scheffé. Les simplexes centrés ajoutent des mélanges supplémentaires dans le centre du domaine d'expérimentation par rapport aux simplexes classiques. Cela a pour conséquence d'améliorer la qualité des prédictions au centre du domaine.

Les réseaux simplexes augmentés définissent un maillage plus important de points. Cela améliore la qualité de prédiction du modèle à l'intérieure du domaine.

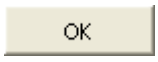
Par défaut XLSTAT associe un modèle canonique réduit (Simplified Canonical Model) aux simplexes centrés. Toutefois, il est possible de changer de modèle sous réserve que le nombre de degrés de liberté soit suffisant (en augmentant le nombre de répétitions d'essais).

Sorties

Cet outil vous permet donc d'obtenir un plan d'expérience afin de lancer vos expérimentations. Il permet aussi d'obtenir des feuilles individuelles associées à chaque expérience qui sont générées sur des feuilles Excel qui pourront être imprimées et remplies.

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.


: cliquez sur ce bouton pour changer la façon dont vous chargez les données dans XLSTAT. Si le bouton possède une icône de souris, XLSTAT vous permet de faire une sélection à la souris des données. Si le bouton possède une icône de liste, XLSTAT vous permet de sélectionner les données à partir d'une liste.

Onglet **Général** :

Facteurs quantitatifs (min/max) : sélectionnez le minimum et le maximum des facteurs quantitatifs. La sélection doit comporter deux lignes, la première correspond aux minimums, la seconde aux maximums pour chacun des facteurs.

Nombre de réponses : entrez le nombre de réponses à analyser.

Répétitions : activez cette option afin de sélectionner le nombre de répétitions du plan d'expériences.

Ordre aléatoire : activez cette option si vous souhaitez que l'ordre des expériences dans le plan soit aléatoire.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des variables : activez cette option si la première ligne de la sélection contient le libellé des variables.

Onglet **Options** :

Plan d'expérience : choisissez le plan d'expérience que vous désirez utiliser. Les plans proposés sont ceux expliqués en introduction : le simplexe, le simplexe centré, et le simplexe centré augmenté.

Nombre de degrés du modèle : dans le cas d'un plan simplexe, vous avez la possibilité de choisir le nombre de degrés du modèle.

Quantité totale du mélange : entrez la quantité totale du mélange. Elle est la quantité du mélange utilisée dans l'expérience.

Onglet **Sorties** :

Afficher les feuilles d'expérience : activez cette option si vous désirez afficher des feuilles Excel individuelles pour chaque expérience. Ceci peut être utile afin de les imprimer et de réaliser les expériences.

Résultats

Information sur les variables : ce tableau regroupe les informations sur les facteurs. Pour chaque facteur, le nom court, le nom long et l'unité utilisée sont affichés.

Plan d'expériences : le plan d'expériences est affiché dans ce tableau. Des colonnes supplémentaires contenant des informations sur les facteurs et sur les réponses, ainsi qu'un libellé pour chaque expérience, l'ordre de tri, l'ordre d'exécution et le numéro de la répétition sont aussi inclus dans ce tableau.

Optimisation des réponses : le tableau d'optimisation des réponses est affiché à la suite du plan d'expériences. Vous devez alors sélectionner les paramètres suivants :

- **Objectif** : choisissez l'objectif de l'optimisation. Vous avez le choix entre minimum, optimum et maximum.

Si l'objectif sélectionné est l'optimum ou le maximum, les champs suivants sont activés :

- **Inférieur** : entrez pour chaque réponse la valeur de la borne inférieure en dessous de laquelle la désirabilité vaut 0.
- **Cible (gauche)** : entrez pour chaque réponse la valeur de la borne inférieure au-dessus de laquelle la désirabilité vaut 1.

Si l'objectif sélectionné est l'optimum ou le minimum, les champs suivants sont activés :

- **Cible (droite)** : entrez pour chaque réponse la valeur de la borne supérieure en dessous de laquelle la désirabilité vaut 1.
- **Inférieur** : entrez pour chaque réponse la valeur de la borne supérieure au-dessus de laquelle la désirabilité vaut 0.
- **s** : activez cette option si la fonction de désirabilité croissante doit être non linéaire. Entrez alors la valeur du paramètre de forme. Cette valeur doit être comprise entre 0,01 et 100.
- **t** : activez cette option si la fonction de désirabilité décroissante doit être non linéaire. Entrez alors la valeur du paramètre de forme. Cette valeur doit être comprise entre 0,01 et 100.
- **Poids** : activez cette option si les réponses doivent avoir une valeur exponentielle différente de 1 lors du calcul de la désirabilité. Entrez alors la valeur du paramètre de

forme. Cette valeur doit être comprise entre 0,01 et 100.

Pour plus de détails sur l'optimisation des réponses, vous pouvez vous référez à l'aide de l'analyse d'un plan d'effet de facteurs

Exemple

Un exemple de génération et d'analyse d'un plan de mélange est disponible sur le Centre d'aide XLSTAT à l'adresse suivante :

http://www.xlstat.com/demo-mixturef.htm

Bibliographie

Droesbeke J.J., Fine J. and Saporta G. (1997). Plans d'Expériences - Application Industrielle. Editions Technip.

Louvet F. and Delplanque L. (2005). Design Of Experiments: The French touch, Les plans d'expériences : une approche pragmatique et illustrée, Alpha Graphic, Olivet.

Scheffé H. (1958). Experiments with mixture. *Journal of Royal Statistical Society*, B, **20**, 344-360.

Scheffé H. (1958). Simplex-centroid design for experiments with mixtures. *Journal of Royal Statistical Society*, B, **25**, 235-263.

Analyse d'un plan de mélange

Utilisez cet outil pour analyser un plan de mélange généré avec la fonction de génération de plans de mélange de XLSTAT.

Dans cette section :

Description

L'analyse d'un plan de mélange se base sur le même principe qu'une régression linéaire. La différence majeure vient du modèle qui est utilisé. Plusieurs modèles sont utilisés.

Par défaut XLSTAT associe un modèle canonique réduit (Simplified Canonical Model) aux simplexes centrés. Toutefois, il est possible de changer de modèle sous réserve que le nombre de degrés de liberté soit suffisant (en augmentant le nombre de répétitions d'essais).

Afin de vérifier la contrainte associée au plan de mélange, on utilise un modèle polynomial sans constante. On différencie deux types de modèles, les modèles réduits et les modèles complets (à partir du degré 3).

Les équations du modèle sont les suivantes :

- Modèle linéaire (degré 1) :

$$Y = \sum_i \beta_i x_i$$

- Modèle quadratique (degré 2) :

$$Y = \sum_i \beta_i x_i + \sum_i \sum_{i<j} \beta_{ij} x_i x_j$$

- Modèle cubique complet (degré 3) :

$$Y = \sum_i \beta_i x_i + \sum_i \sum_{i<j} \beta_{ij} x_i x_j + \sum_j \sum_{i<j} \delta_{ij} x_i x_j (x_i - x_j) + \sum_k \sum_{j<k} \beta_{ijk} x_i x_j (x_k)$$

- Modèle cubique simplifié (spécial) :

$$Y = \sum_i \beta_i x_i + \sum_i \sum_{i<j} \beta_{ij} x_i x_j + \sum_k \sum_{j<k} \beta_{ijk} x_i x_j (x_k)$$

XLSTAT permet d'aller jusqu'au degré 4 (modèle quartique simplifié et complet).

L'estimation de ces modèles se fait par régression classique. Pour plus de détails sur l'ANOVA et la régression linéaire, veuillez vous référer aux chapitres de l'aide à ce sujet.

Optimisation des réponses et désirabilité

Il est possible d'optimiser chaque réponse de manière individuelle et de combiner les résultats afin d'obtenir une fonction de désirabilité et d'analyser ses valeurs. Introduite par Derringer and Suich (1980), cette approche est basée sur la transformation de la réponse y_i en une fonction de désirabilité individuelle d_i sur l'intervalle $0 \leq d_i \leq 1$.

Lorsque y_i a atteint sa valeur cible, alors $d_i égal 1$. Si y_i se trouve en dehors d'une région acceptable autour de la cible, alors $d_i = 0$. Entre ces deux cas, des valeurs intermédiaires de d_i sont obtenues.

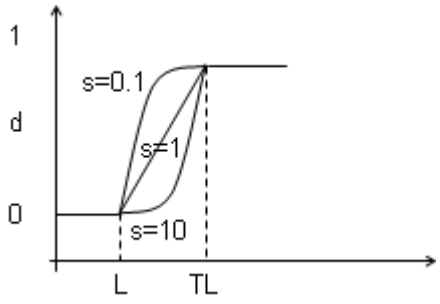
Les 3 différents cas pour d_i sont :

- L = borne inférieure. Pour chaque valeur $< L$, on a $d_i = 0$
- U = borne supérieure. Pour chaque valeur $> U$, on a $d_i = 0$.
- $T(L)$ = valeur cible à gauche.
- $T(R)$ = valeur cible à droite. Pour chaque valeur entre $T(L)$ et $T(R)$, on a $d_i = 1$.
- s, t = paramètres permettant de définir la forme de la fonction d'optimisation entre L et $T(L)$ et entre $T(R)$ et U.

L'équation suivante doit être respectée lorsqu'on définit L, U, $T(L)$ et $T(R)$:

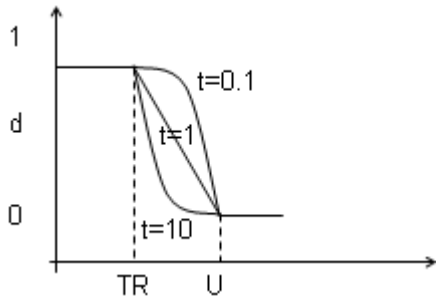
$$L \leq T(L) \leq T(R) \leq U$$

Maximiser la valeur de y_i :



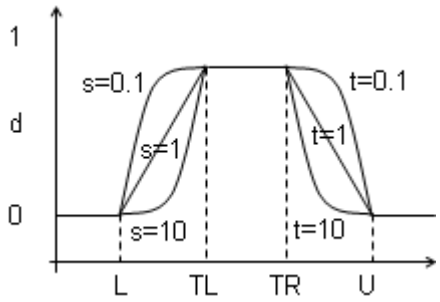
$$d_i = \begin{cases} 0 & y_i < L \\ (\frac{y_i - L}{T(L) - L})^s & L \leq y_i \leq T(L) \\ 1 & y_i > T(L) \end{cases}$$

Minimiser la valeur de y_i :



$$d_i = \begin{cases} 1 & y_i < T(R) \\ (\frac{U-y_i}{U-T(R)})^t & T(R) \leq y_i \leq U \\ 0 & y_i > U \end{cases}$$

Afin de cibler un intervalle donné de y_i, on peut utiliser la fonction de désirabilité suivante :



$$d_i = \begin{cases} 0 & y_i < L \\ (\frac{y_i-L}{T(L)-L})^s & L \leq y_i \leq T(L) \\ 1 & T(L) < y_i < T(R) \\ (\frac{U-y_i}{U-T(R)})^t & T(R) \leq y_i \leq U \\ 0 & y_i > U \end{cases}$$

Les paramètres sont choisis de manière à maximiser la désirabilité globale D.

$$D = (d_1^{w_1} \cdot d_2^{w_2} \cdot ... \cdot d_m^{w_m})^{\frac{1}{w_1 \cdot w_2 \cdot ... \cdot w_m}}$$

Où $1 \leq w_i \leq 10$ sont des poids associés aux fonctions de désirabilité individuelles. Plus les w_i sont grands, plus les d_i sont pris en compte lors de l'optimisation.

Au moment de l'affichage, XLSTAT donne les 5 meilleures solutions trouvées lors de l'optimisation.

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.

OK : cliquez sur ce bouton pour lancer les calculs.

Annuler : cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

Aide : cliquez sur ce bouton pour afficher l'aide.

 : cliquez sur ce bouton pour rétablir les options par défaut.

 : cliquez sur ce bouton pour effacer les sélections de données.

 : cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.

Onglet **Général** :

Y / résultats : sélectionnez les colonnes du plan d'expériences correspondant aux réponses. Ces colonnes doivent maintenant être remplies avec les résultats des expériences qui ont été menées.

Plan d'expériences : sélectionnez votre plan d'expériences. Si vous avez modifié votre plan, vérifiez que les facteurs qualitatifs et quantitatifs se suivent. L'ensemble des colonnes du plan doivent être sélectionnées.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des variables : activez cette option si la première ligne des données sélectionnées contient un libellé.

Onglet **Réponses** :

Optimisation des réponses : activez cette option si vous souhaitez réaliser une optimisation des réponses. Dans ce cas sélectionnez le tableau d'optimisation des réponses généré au moment de la création du plan. L'en-tête du tableau doit être inclus dans la sélection.

- **Objectif** : choisissez l'objectif de l'optimisation. Vous avez le choix entre minimum, optimum et maximum.

Si l'objectif sélectionné est l'optimum ou le maximum, les champs suivants sont activés :

- **Inférieur** : entrez pour chaque réponse la valeur de la borne inférieure en dessous de laquelle la désirabilité vaut 0.

- **Cible (gauche)** : entrez pour chaque réponse la valeur de la borne inférieure au-dessus de laquelle la désirabilité vaut 1.

Si l'objectif sélectionné est l'optimum ou le minimum, les champs suivants sont activés :

- **Cible (droite)** : entrez pour chaque réponse la valeur de la borne supérieure en dessous de laquelle la désirabilité vaut 1.
- **Inférieur** : entrez pour chaque réponse la valeur de la borne supérieure au-dessus de laquelle la désirabilité vaut 0.
- **s** : activez cette option si la fonction de désirabilité croissante doit être non linéaire. Entrez alors la valeur du paramètre de forme. Cette valeur doit être comprise entre 0,01 et 100.
- **t** : activez cette option si la fonction de désirabilité décroissante doit être non linéaire. Entrez alors la valeur du paramètre de forme. Cette valeur doit être comprise entre 0,01 et 100.
- **Poids** : activez cette option si les réponses doivent avoir une valeur exponentielle différente de 1 lors du calcul de la désirabilité. Entrez alors la valeur du paramètre de forme. Cette valeur doit être comprise entre 0,01 et 100.

Modèle : sélectionnez le type de modèle que vous souhaitez utiliser lors de votre analyse. En fonction du modèle choisi, les facteurs correspondant à ce modèle seront présélectionnés lors de l'affichage de la seconde interface. Vous pourrez alors sélectionner ou désélectionner les facteurs que vous souhaitez utiliser.

Quantité totale du mélange : entrez la quantité totale du mélange. Elle est la quantité du mélange utilisée dans l'expérience. Elle doit correspondre à celle entrée lors de la création du plan de mélange.

Onglet **Sorties** :

Corrélations : activez cette option pour afficher la matrice de corrélation pour les variables quantitatives (dépendantes et explicatives).

Coefficients d'ajustement : activez cette option pour afficher le tableau des indices de qualité d'ajustement du modèle.

Analyse de la variance : activez cette option pour afficher le tableau d'analyse de la variance.

Contributions : activez cette option pour afficher le tableau des contributions. Cette option est nécessaire si vous souhaitez afficher le diagramme de Pareto.

Coefficients normalisés : activez cette option pour afficher les paramètres normalisés du modèle (coefficients bêta).

Prédictions et résidus : activez cette option pour afficher les prédictions et les résidus pour l'ensemble des observations.

- **Prédictions ajustées** : activez cette option pour calculer et afficher les prédictions ajustées dans le tableau des prédictions et résidus.

- **D de Cook** : activez cette option pour calculer et afficher les distances de Cook dans le tableau des prédictions et résidus.
- **Résidus studentisés** : activez cette option pour calculer et afficher les résidus studentisés dans le tableau des prédictions et résidus.

Onglet **Graphiques** :

Graphiques de régression : activez cette option pour afficher les graphiques de régression :

- **Coefficients normalisés** : activez cette option pour afficher sur un graphique les paramètres normalisés du modèle avec leur intervalle de confiance.
- **Prédictions et résidus** : activez cette option pour afficher les graphiques suivants :

(1) Droite de régression : ce graphique n'est affiché que s'il n'y a qu'une seule variable explicative, et que cette variable est quantitative.

(2) Variable explicative versus résidus normalisés : ce graphique n'est affiché que s'il n'y a qu'une seule variable explicative, et que cette variable est quantitative.

(3) Variable dépendante versus résidus normalisés.

(4) Prédictions pour la variable dépendante versus variable dépendante.

(5) Graphique en bâtons des résidus normalisés.

Intervalles de confiance : activez cette option pour afficher les intervalles de confiance sur les graphiques (1) et (4).

Graphe ternaire : activez cette option pour afficher un graphique ternaire. Ce graphique s'affiche à partir de 3 facteurs.

Résultats

Information sur les variables : ce tableau récapitule les informations sur les facteurs. Pour chaque facteur, le nom court, le nom long et l'unité utilisée sont affichés.

Optimisation des réponses : ce tableau donne les 5 meilleures solutions obtenues lors de l'optimisation des réponses.

Matrice de corrélation : ce tableau est affiché afin de vous permettre d'avoir un aperçu des corrélations entre les différentes variables sélectionnées.

Coefficients d'ajustement : dans ce tableau sont affichées les statistiques relatives à l'ajustement du modèle de régression :

- **Observations** : le nombre d'observations prises en compte dans les calculs. Dans les formules présentées ci-dessous n désigne le nombre d'observations.
- **Somme des poids** : la somme des poids des observations prises en compte dans les calculs. Dans les formules présentées ci-dessous W désigne la somme des poids.

- **DDL** : le nombre de degrés de liberté pour le modèle retenu (correspondant à la partie erreurs).

- **R²** : le coefficient de détermination du modèle. Ce coefficient, dont la valeur est comprise entre 0 et 1, n'est affiché que si la constante du modèle n'est pas fixée par l'utilisateur. Sa valeur est définie par

$$R^2 = 1 - \frac{\sum_{i=1}^{n} w_i (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} w_i (y_i - \bar{y})^2} \text{ avec } \bar{y} = \frac{1}{n} \sum_{i=1}^{n} w_i y_i$$

- Le R^2 s'interprète comme la proportion de la variabilité de la variable dépendante expliquée par le modèle. Plus le R^2 est proche de 1, meilleur est le modèle. L'inconvénient du R^2 est qu'il ne prend pas en compte le nombre de variables utilisées pour ajuster le modèle.

- **R²ajusté** : le coefficient de détermination ajusté du modèle. Le R^2 ajusté peut être négatif si le R^2 est voisin de zéro. Ce coefficient n'est affiché que si la constante du modèle n'est pas fixée par l'utilisateur. Sa valeur est définie par

$$\hat{R}^2 = 1 - (1 - R^2) \frac{W - 1}{W - p - 1}$$

Le R^2 ajusté est une correction du R^2 qui permet de prendre en compte le nombre de variables utilisées dans le modèle.

- **MCE** : la moyenne des carrés des erreurs (MCE) est définie par :

$$MCE = \frac{1}{W - p^*} \sum_{i=1}^{n} w_i (y_i - \hat{y}_i)^2$$

- **RMCE** : la racine de la moyenne des carrés des erreurs (RMCE) est la racine carrée de la MCE.

- **MAPE** : la *Mean Absolute Percentage Error* est calculée comme suit :

$$MAPE = \frac{1}{W - p^*} \sum_{i=1}^{n} w_i \left| \frac{y_i - \hat{y}_i}{y_i} \right|$$

- **DW** : le coefficient de Durbin-Watson est défini par

$$DW = \frac{\sum_{i=2}^{n} [(y_i - \hat{y}_i) - (y_{i-1} - \hat{y}_{i-1})]^2}{\sum_{i=1}^{n} w_i (y_i - \hat{y}_i)^2}$$

Ce coefficient correspond au coefficient d'autocorrélation d'ordre 1 et permet de vérifier que les résidus du modèle ne sont pas autocorrélés, sachant que l'indépendance des résidus est l'une

des hypothèses de base de la régression linéaire. L'utilisateur pourra se référer à une table des coefficients de Durbin-Watson pour vérifier si l'hypothèse d'indépendance des résidus est acceptable.

- **Cp** : le coefficient Cp de Mallows est défini par

$$Cp = \frac{SCE}{\hat{\sigma}} + 2p^* - W$$

où SCE est la somme du carré des erreurs pour le modèle avec p variables explicatives, et où $\hat{\sigma}$ correspond à l'estimateur de la variance des résidus pour le modèle comprenant toutes les variables explicatives. Plus le coefficient Cp est proche de p^* moins le modèle est biaisé.

- **AIC** : le critère d'information d'Akaike (Akaike's Information Criterion) est défini par

$$AIC = W \ln(\frac{SCE}{W}) + 2p^*$$

Ce critère proposé par Akaike (1973) dérive de la théorie de l'information, et s'appuie sur la mesure de Kullback et Leibler (1951). C'est un critère de sélection de modèles qui pénalise les modèles pour lesquels l'ajout de nouvelles variables explicatives n'apporte pas suffisamment d'information au modèle, l'information étant mesurée au travers de la SCE. On cherche à minimiser le critère AIC.

- **SBC** : le critère bayésien de Schwarz (Schwarz's Bayesian Criterion) est défini par

$$SBC = W \ln(\frac{SCE}{W}) + \ln(W)p^*$$

Ce critère proposé par Schwarz (1978) est proche du critère AIC, et comme ce dernier on cherche à le minimiser.

- **PC** : le critère de prédiction d'Amemiya (Amemiya's Prediction Criterion) est défini par

$$PC = \frac{(1-R^2)(W+p^*)}{W-p^*}$$

Ce critère proposé par Amemiya (1980) permet comme le R^2 ajusté de tenir compte de la parcimonie du modèle.

- **Press** : la statistique du Press (predicted residual error sum of squares) n'est affichée que si l'option correspondante a été activée dans la boîte de dialogue. Elle est définie par

$$Press = \sum_{i=1}^{n} w_i (y_i - \hat{y}_{i(-i)})^2$$

où $\hat{y}_{i(-i)}$ est la prédiction pour l'observation i lorsque cette dernière n'est pas utilisée pour l'estimation des paramètres. On obtient alors

$$\text{Press RMCE } = \sqrt{\frac{\text{Press}}{W-p^*}}$$

Le Press RMCE peut alors être comparé au RMCE. Une différence importante entre les deux indique que le modèle est sensible à la présence ou absence de certaines observations dans le modèle.

- **Q²** : La statistique Q^2 est affichée. Elle est définie par :

$$Q^2 = 1 - \frac{PressRMSE}{SSE}$$

Plus cet indice est proche de 1, plus le modèle est bon et robuste.

Le **tableau d'analyse de la variance** permet d'évaluer le pouvoir explicatif des variables explicatives. Dans le cas où la constante du modèle n'est pas fixée à une valeur donnée, le pouvoir explicatif est évalué en comparant l'ajustement (au sens des moindres carrés) du modèle final avec l'ajustement du modèle rudimentaire composé d'une constante égale à la moyenne de la variable dépendante. Dans le cas où la constante du modèle est fixée, la comparaison est faite par rapport au modèle pour lequel la variable dépendante serait égale à la constante fixée.

Si l'option **Type I/III SS** (SS : Sum of Squares) est activée, les tableaux suivants sont affichés.

Le tableau des **Type I SS** permet de visualiser l'influence de l'ajout progressif des variables explicatives sur l'ajustement du modèle, au sens de la somme des carrés des erreurs (SCE), de la moyenne des carrés des erreurs (MCE), du F de Fisher, ou de la probabilité associée au F de Fisher. Plus la probabilité est faible, plus la contribution de la variable au modèle est importante, toutes les autres variables étant déjà dans le modèle. Remarques : l'ordre de sélection des variables dans le modèle influe sur les valeurs obtenues ; la somme des sommes des carrés de ce tableau est égale à la somme des carrés du modèle.

Le tableau des **Type II SS** permet de visualiser l'influence du retrait d'une variable explicative sur l'ajustement du modèle, toutes les autres variables étant conservées, au sens de la somme des carrés des erreurs (SCE), de la moyenne des carrés des erreurs (MCE), du F de Fisher, ou de la probabilité associée au F de Fisher. Plus la probabilité est faible, plus la contribution de la variable au modèle est importante. Remarque : dans le cas des ANOVAs déséquilibrées, l'utilisation des Type III est recommandée mais XLSTAT affiche les Type II pour les utilisateurs avancés qui voudraient disposer des Type II.

Le tableau des **Type III SS** permet de visualiser l'influence du retrait d'une variable explicative sur l'ajustement du modèle, toutes les autres variables étant conservées, au sens de la somme des carrés des erreurs (SCE), de la moyenne des carrés des erreurs (MCE), du F de Fisher, ou de la probabilité associée au F de Fisher. Plus la probabilité est faible, plus la contribution de la variable au modèle est importante, toutes les autres variables étant déjà dans le modèle. Remarque : contrairement au cas des Type I SS, l'ordre de sélection des variables dans le modèle n'influe pas sur les valeurs obtenues, et contrairement aux Type II SS, les valeurs ne dépendent pas des effectifs des cellules (par cellule on entend une combinaison de modalités des différents facteurs), ce qui fait des Type III le test recommandé pour évaluer la contribution d'une variable.

Le tableau **paramètres du modèle** affiche l'estimation des paramètres, l'erreur type correspondante, le t de Student, la probabilité correspondante, ainsi que l'intervalle de confiance.

L'**équation du modèle** est ensuite affichée pour faciliter la lecture ou la réutilisation du modèle.

Le tableau des **coefficients normalisés** (aussi appelés coefficients bêta) permet de comparer le poids relatif des variables. Plus la valeur absolue d'un coefficient est élevée, plus le poids de la variable correspondante est important. Lorsque l'intervalle de confiance autour des coefficients normalisés comprend la valeur 0 (cela est facilement visible sur le graphique des coefficients normalisés), le poids d'une variable dans le modèle n'est pas significatif.

Dans le tableau des **prédictions et résidus** sont donnés pour chaque observation, son poids, la valeur de la variable explicative qualitative s'il n'y en a qu'une, la valeur observée de la variable dépendante, la prédiction du modèle, les résidus, les intervalles de confiance, ainsi que la prédiction ajustée et le D de Cook si les options correspondantes ont été activées dans la boîte de dialogue. Deux types d'intervalles de confiance sont affichés : un intervalle de confiance autour de la moyenne (correspondant au cas où l'on ferait la prédiction pour un nombre infini d'observations avec un ensemble de valeurs données des variables explicatives) et un intervalle autour de la prédiction ponctuelle (correspondant au cas d'une prédiction isolée pour des valeurs données des variables explicatives). Le second intervalle est toujours plus grand que le premier, les aléas étant plus importants. Si des données de validation ont été sélectionnées, elles sont affichées en fin de tableau.

Les **graphiques** qui suivent permettent de visualiser les résultats mentionnés ci-dessus. S'il n'y a qu'une seule variable explicative dans le modèle, le premier graphique affiché permet de visualiser les valeurs observées, la droite de régression et les deux types d'intervalles de confiance autour des prévisions. Le second graphique permet quant à lui de visualiser les résidus normalisés en fonction de la variable explicative. En principe, les résidus doivent être distribués de manière aléatoire autour de l'axe des abscisses. L'observation d'une tendance ou d'une forme révèlerait un problème au niveau du modèle.

Les **trois graphiques** affichés ensuite permettent de visualiser respectivement l'évolution des résidus normalisés en fonction de la variable dépendante, la distance entre les prédictions et les observations (pour un modèle idéal, les points seraient tous sur la bissectrice), et les résidus normalisés sous la forme d'un diagramme en bâtons. Ce dernier graphique permet de rapidement voir si un nombre anormal de données sort de l'intervalle $]-2,2[$ sachant que ce dernier, sous hypothèse de normalité, doit contenir environ 95% des données.

Graphe ternaire : pour chaque combinaison de facteurs, on trace un graphe ternaire. Ce graphe représente une surface de réponse sur une des faces du polyèdre que constitue le domaine expérimental. Ces graphiques facilitent l'interprétation du modèle et permettent d'identifier les configurations optimales.

Exemple

Un exemple de génération et d'analyse d'un plan de mélange est disponible sur le Centre d'aide XLSTAT à l'adresse suivante :

http://www.xlstat.com/demo-mixturef.htm

Bibliographie

Droesbeke J.J., Fine J. and Saporta G. (1997). Plans d'Expériences - Application Industrielle. Editions Technip.

Louvet F. and Delplanque L. (2005). Design Of Experiments: The French touch, Les plans d'expériences : une approche pragmatique et illustrée, Alpha Graphic, Olivet.

Scheffé H. (1958). Experiments with mixture. *Journal of Royal Statistical Society*, B, **20**, 344-360.

Scheffé H. (1958). Simplex-centroid design for experiments with mixtures. *Journal of Royal Statistical Society*, B, **25**, 235-263.

Plans de Taguchi

Utilisez cet outil pour générer des plans de Taguchi dans le but de trouver des produits robustes et insensibles à la variabilité naturelle de l'environnement et des processus.

Dans cette section :

Description

La méthode de Taguchi est une méthode introduite par Genichi Taguchi (Genichi et Wu, 1980) qui est une méthode de plans d'expériences apportant une amélioration aux plans factoriels complets et fractionnaires.

Cette méthode nécessite deux étapes :

- Une étape de génération du plan de Taguchi durant laquelle il est possible de choisir parmi une liste de plans en fonction du nombre de facteurs et des niveaux de ceux-ci.
- Une fois le plan de Taguchi obtenu, vous pouvez l'analyser afin d'identifier les paramètres de facteur de contrôle qui minimisent la variation de la réponse. Cette seconde étape est réalisable avec l'outil Analyse d'un plan de Taguchi.

La méthode de Taguchi

L'approche de G. Taguchi est une des techniques d'ingénierie la plus utilisée dans le monde. Cette méthode vise à développer des produits ayant un bon fonctionnement malgré les variations naturelles des matériaux, des opérateurs, de l'environnement.

Elle consiste à utiliser des tables orthogonales, qui ont été préétablies par G. Taguchi, qui dépendent du nombre d'essais à réaliser, du nombre de facteurs composant le modèle, et du nombre de niveaux par facteurs.

La méthode de Taguchi divise les problèmes d'optimisation en deux catégories : la méthode statique et la méthode dynamique.

Les plans de Taguchi statiques ont pour but de déterminer les meilleurs niveaux de facteurs de contrôle afin que la sortie soit à la valeur souhaitée.

Les plans dynamiques ont, quant à eux, un facteur signal. L'objectif est de déterminer les meilleurs niveaux de facteurs de contrôle afin d'améliorer la relation entre ce facteur de signal et une réponse de sortie.

En fonction du nombre de facteurs et de leurs niveaux, XLSTAT propose, dans une nouvelle boite de dialogue, la liste des plans qu'il est possible de réaliser. Ces plans ont une notation particulière définie comme cela : L(nbEssais)(nbNiveaux^nbFacteurs)

Où nbEssais = nombre d'essais, nbNiveaux = nombre de niveaux pour chaque facteur, et nbFacteurs = nombre de facteurs.

Si la notation est de cette forme : L(nbEssais)(nbNiveaux^nbFacteurs nbNiveaux^nbFacteurs), le plan contient des facteurs à différents niveaux.

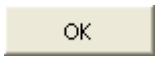
Par exemple, un plan L9(3^3) est un plan ayant 9 essais et 3 facteurs à 3 niveaux. Un plan L18(3^3 6^1) est un plan ayant 18 essais, 3 facteurs à 3 niveaux et 1 facteur à 6 niveaux.

La liste des plans disponibles dans XLSTAT est la suivante :

- $L4(2^3)$
- $L8(2^7)$
- $L9(3^4)$
- $L12(2^{11})$
- $L16(2^{15})$
- $L16(4^5)$
- $L25(5^6)$
- $L27(3^{13})$
- $L32(2^{31})$
- $L32(2^1 4^9)$
- $L36(2^{11} 3^{12})$
- $L36(2^3 3^{13})$
- $L50(2^1 4^{11})$
- $L8(4^4 2^1)$
- $L16(4^1 2^{12})$
- $L16(4^2 2^9)$
- $L16(4^3 2^6)$
- $L16(4^4 2^3)$
- $L18(2^1 3^7)$
- $L18(3^6 6^1)$

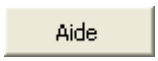
Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.



: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.

Onglet **Général** :

Tableau facteurs/modalités : entrez le tableau regroupant le nom des facteurs et leurs modalités.

Nombre de réponses : entrez le nombre de réponses de votre analyse.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des variables : activez cette option si la première ligne des données sélectionnées contient un libellé.

Onglet **Options** :

Ajouter un facteur de signal : activez cette option pour ajouter un facteur signal au plan de Taguchi.

- **Nombre de niveaux** : sélectionnez le nombre de niveaux du facteur signal.
- **Labels du signal** : sélectionnez les labels de chacun des niveaux du facteur signal. Ces labels doivent être numériques.

Interactions : activez cette option pour ajouter des interactions au plan de Taguchi.

Onglet **Sorties** :

Plan codé : activez cette option si vous souhaitez afficher le plan codé, c'est à dire le plan de Taguchi contenant des 1, 2, 3, ... plutôt que les catégories de vos facteurs.

Résultats

Informations sur les variables : dans ce tableau sont récapitulées toutes les informations sur les facteurs sélectionnés.

Plan d'expérience : ce tableau affiche le plan de Taguchi. Si un facteur signal est ajouter, une colonne supplémentaire apparait dans le tableau du plan d'expérience. Les colonnes vides servent à être remplies avec les réponses.

Lancer l'analyse : une fois que les colonnes réponses sont remplies, vous pouvez cliquer sur le bouton « Lancer l'analyse » afin d'ouvrir la boite de dialogue préremplie permettant d'effectuer l'analyse du plan de Taguchi.

Plan codé : ce tableau affiche le plan de Taguchi codé, contenant des 1, 2, 3, ... plutôt que les catégories de vos facteurs.

Exemple

Un exemple d'analyse d'un plan de Taguchi est disponible sur le Centre d'aide XLSTAT :

http://www.xlstat.com/demo-taguchif.htm

Bibliographie

Taguchi, G., & Wu, Y. (1980). Introduction to Off-Line Quality Control. Central Japan Quality Control Association.

Taguchi, G., Chowdhury, S., Wu, Y. & Yano, H. (2005). Taguchi's Quality Engineering Handbook. Hoboken, N.J., John Wiley & Sons.

Sabre, R. (2007). Plans d'expériences - Méthode de Taguchi. Techn. l'Ingénieur, Tech. Rep. F1006 V1.

Analyse d'un plan de Taguchi

Utilisez cet outil pour analyser un plan de Taguchi dans le but de trouver des produits robustes et insensibles à la variabilité naturelle de l'environnement et des processus.

Dans cette section :

Description

La méthode de Taguchi est une méthode introduite par Genichi Taguchi (Taguchi et Wu, 1980) qui est une méthode de plans d'expériences apportant une amélioration aux plans factoriels complets et fractionnaires.

Elle consiste à utiliser des tables orthogonales, qui ont été préétablies par G. Taguchi, ces tables dépendent du nombre d'essais à réaliser, du nombre de facteurs composant le modèle, et du nombre de niveaux par facteurs.

Dans les plans d'expériences classiques, l'objectif est d'identifier les facteurs qui affectent la réponse moyenne et de les contrôler aux niveaux que l'on souhaite. Les plans d'expériences de Taguchi traitent conjointement la moyenne et la variabilité des valeurs des caractéristiques mesurées grâce à l'utilisation des rapports signal sur bruit (S/B).

La méthode de Taguchi

L'approche de G. Taguchi est une des techniques d'ingénierie la plus utilisée dans le monde. Cette méthode vise à développer des produits ayant un bon fonctionnement malgré les variations naturelles des matériaux, des opérateurs, de l'environnement.

Dans le but de trouver des produits robustes et insensibles à la variabilité, XLSTAT propose d'étudier trois paramètres : le rapport signal sur bruit, la moyenne (ou la pente) et l'écart-type.

XLSTAT permet aussi d'ajuster le modèle linéaire pour ces trois paramètres. Un tableau des coefficients de regression estimés sera alors affiché, avec lequel il sera possible de déterminer quels sont les facteurs ayant des valeurs statistiquement significatives au seuil $\alpha = 0.05$.

Le rapport signal sur bruit

Le rapport signal sur bruit est différent selon si on utilise un plan de Taguchi statique, ou un plan dynamique.

Rapport signal sur bruit statique :

Traditionnellement, une seule sortie d'un produit ou d'un processus est utilisée en recherche et développement. Dans ce cas, on utilise un rapport signal sur bruit statique pour améliorer la robustesse du produit en question. Pour un rapport signal sur bruit statique, il existe deux problèmes : réduire la variabilité, et ajuster la moyenne.

En fonction de l'objectif de votre expérience, plusieurs rapports signal sur bruit sont disponibles dans XLSTAT :

- **Préférer plus grand** : sélectionnez cette option si l'intention est de maximiser la réponse et qu'il n'y a pas de données négatives. Pour calculer ce rapport on utilise la formule suivante :

$$S/B = -10 \times \log\left(\frac{\sum(\frac{1}{Y^2})}{n}\right)$$

où Y est la réponse pour la combinaison de niveaux de facteurs donnée et n est le nombre de réponses dans la combinaison de niveaux de facteurs.

- **Préférer nominal : type I** : sélectionnez cette option si il n'y a pas de données négatives. Pour une application "Préférer nominal", on réalise une optimisation en deux étapes. La première consiste à maximiser le rapport signal sur bruit, la seconde à ajuster la moyenne. Pour calculer ce rapport on utilise la formule suivante :

$$S/B = 10 \times \log\left(\frac{\bar{Y}^2}{s^2}\right)$$

où $\bar{Y}$ est la moyenne des réponses pour la combinaison de niveaux de facteurs donnée, s est l'écart-type des réponses pour la combinaison de facteurs donnée et n est le nombre de réponses dans la combinaison de niveaux de facteurs.

- **Préférer nominal : type II** : sélectionnez cette option lorsque les résultats d'une expérience incluent des valeurs négatives. Ainsi, les valeurs positives et négatives s'annulent. Les informations concernant la sensibilité ou la moyenne ne peuvent pas être obtenues. On choisi alors ce type de rapport signal sur bruit, qui ne montre que la variabilité, il est cependant moins informatif que le type I. Pour calculer ce rapport on utilise la formule suivante :

$$S/B = -10 \times \log(s^2)$$

où s est l'écart-type des réponses pour tous les facteurs de bruit de la combinaison de niveaux de facteurs donnée.

- **Préferer plus petit** : sélectionnez cette option si l'intention est de minimiser la réponse et qu'il n'y a pas de données négatives. Pour calculer ce rapport on utilise la formule suivante :

$$S/B = -10 \times \log\left(\frac{\sum Y^2}{n}\right)$$

où Y est la réponse pour la combinaison de niveaux de facteurs donnée et n est le nombre de réponses dans la combinaison de niveaux de facteurs.

Rapport signal sur bruit dynamique :

Dans le cas d'un plan de Taguchi dynamique, XLSTAT utilise une équation proportionnelle au point zéro. C'est à dire que l'entrée est égale à zéro, et que la sortie doit passer par l'origine. Pour calculer ce rapport on utilise la formule suivante :

$$S/B = 10 \times \log\left(\frac{pente^2}{MCE}\right)$$

où la pente est le taux de variation de la réponse par rapport au facteur de signal, et MCE est la Moyenne des Carrés des erreurs.

La moyenne

La moyenne est utilisée uniquement dans le cas de plans de Taguchi statiques. Elle est la réponse moyenne pour chaque combinaison de niveaux de facteurs de contrôle. En fonction de l'objectif de l'expérience on va chercher à minimiser ou maximiser ces moyennes.

La pente

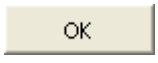
Dans le cas d'un plan de Taguchi dynamique, on utilise la pente plutôt que la moyenne. La pente est le taux de variation de la réponse par rapport au facteur de signal. Elle est ajustée par la méthode des moindres carrés.

L'écart-type

L'écart-type est la variabilité dans la réponse due au bruit. Le but est de minimiser la variabilité, on va donc chercher à minimiser les écarts-types.

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.

 : cliquez sur ce bouton pour effacer les sélections de données.

 : cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.

Onglet **Général** :

Réponses : sélectionnez les réponses pour les différents niveaux des facteurs de bruit.

Plan de Taguchi : sélectionnez les colonnes du plan de Taguchi correspondant aux différents facteurs de contrôle.

Facteur de signal : activez cette option si vous avez sélectionné un facteur de signal au moment de la création du plan de Taguchi. Sélectionnez alors la colonne correspondant au facteur de signal.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des variables : activez cette option si la première ligne des données sélectionnées contient un libellé.

Onglet **Options** :

Rapports Signal sur Bruit : activez le type de rapports signal sur bruit que vous souhaitez calculer parmi les 4 suivants :

- **Préférer plus grand**
- **Préférer nominal : Type II**
- **Préférer nominal : Type I**
- **Préférer plus petit**

Ajuster le modèle linéaire : activez cette option si vous souhaitez ajuster le modèle linéaire pour les coefficients sélectionnés dans l'onglet **Sorties**.

- **Interactions** : activez cette option si vous souhaitez ajouter des interactions dans le modèle. Afin d'avoir des résultats cohérents, il faut avoir sélectionné les interactions au moment de la création du plan.

Onglet **Sorties** :

Rapports Signal sur Bruit : activez cette option si vous souhaitez afficher le tableau de réponses pour les rapports signal sur bruit.

Moyennes : activez cette option si vous souhaitez afficher le tableau de réponses pour les moyennes. Cette option n'est disponible que dans le cas d'un plan de Taguchi statique.

Pente : activez cette option si vous souhaitez afficher le tableau de réponses pour les pentes. Cette option n'est disponible que dans le cas d'un plan de Taguchi dynamique.

Ecarts-types : activez cette option si vous souhaitez afficher le tableau de réponses pour les écarts-types.

Onglet **Graphiques** :

Rapports Signal sur Bruit : activez cette option si vous souhaitez afficher le graphique des effets principaux pour les rapports signal sur bruit.

Moyennes : activez cette option si vous souhaitez afficher le graphique des effets principaux pour les moyennes. Cette option n'est disponible que dans le cas d'un plan de Taguchi statique.

Pente : activez cette option si vous souhaitez afficher le graphique des effets principaux pour les pentes. Cette option n'est disponible que dans le cas d'un plan de Taguchi dynamique.

Ecarts-types : activez cette option si vous souhaitez afficher le graphique des effets principaux pour les écarts-types.

Résultats

Informations sur les variables : dans ce tableau sont récapitulées toutes les informations sur les attributs sélectionnés.

Rapport Signal sur Bruit : ce tableau affiche les réponses pour les rapports signal sur bruit.

Moyennes : ce tableau affiche les réponses pour les moyennes.

Pentes : ce tableau affiche les réponses pour les pentes.

Ecarts-types : ce tableau affiche les réponses pour les écarts-types.

Régression de la variable :

Coefficients d'ajustement : dans ce tableau sont affichées les statistiques relatives à l'ajustement du modèle d'ANOVA :

- **Observations** : le nombre d'observations prises en compte dans les calculs. Dans les formules présentées ci-dessous n désigne le nombre d'observations.

- **Somme des poids** : la somme des poids des observations prises en compte dans les calculs. Dans les formules présentées ci-dessous W désigne la somme des poids.

- **DDL** : le nombre de degrés de liberté pour le modèle retenu (correspondant à la partie erreurs).

- **R²** : le coefficient de détermination du modèle. Sa valeur est comprise entre 0 et 1. Il est défini par :

$$R^2 = \frac{\sum_{i=1}^{n} w_i(\hat{y}_i - \bar{y}_i)^2}{\sum_{i=1}^{n} w_i(y_i - \bar{y}_i)^2} \text{ avec } \bar{y} = \frac{1}{W}\sum_{i=1}^{n} w_i y_i$$

Le coefficient R^2 s'interprète comme la proportion de la variabilité de la variable dépendante expliquée par le modèle. Plus le coefficient R^2 est proche de 1, meilleur est le modèle. L'inconvénient du R^2 est qu'il ne prend pas en compte le nombre de variables utilisées pour ajuster le modèle.

- **R² ajusté** : le coefficient de détermination ajusté du modèle. Le R^2 ajusté peut être négatif si le R^2 est proche de zéro. Il est défini par :

$$\hat{R}^2 = 1 - (1 - R^2)\frac{W - 1}{W - p - 1}$$

Le R^2 ajusté est une correction du R^2 qui permet de prendre en compte le nombre de variables utilisées dans le modèle.

- **MCE** : la moyenne des carrés des erreurs (MCE) est définie par :

$$MCE = \frac{1}{W - p^*}\sum_{i=1}^{n} w_i(y_i - \hat{y}_i)^2$$

- **RMCE** : la racine de la moyenne des carrés des erreurs (RMCE) est la racine carrée de la MCE.

- **MAPE** : la *Mean Absolute Percentage Error* est calculée comme suit :

$$MAPE = \frac{100}{W}\sum_{i=1}^{n} w_i \left|\frac{y_i - \hat{y}_i}{y_i}\right|$$

- **DW** : le coefficient de Durbin-Watson est défini par

$$DW = \frac{\sum_{i=2}^{n}[(y_i - \hat{y}_i) - (y_{i-1} - \hat{y}_{i-1})]^2}{\sum_{i=1}^{n} w_i(y_i - \hat{y}_i)^2}$$

Ce coefficient correspond au coefficient d'autocorrélation d'ordre 1 et permet de vérifier que les résidus du modèle ne sont pas autocorrélés, sachant que l'indépendance des résidus est l'une des hypothèses de base de la régression linéaire. L'utilisateur pourra se référer à une table des coefficients de Durbin-Watson pour vérifier si l'hypothèse d'indépendance des résidus est acceptable.

- **Cp** : le coefficient Cp de Mallows est défini par

$$Cp = \frac{SCE}{\hat{\sigma}} + 2p^* - W$$

où SCE est la somme du carré des erreurs pour le modèle avec p variables explicatives, et où $\hat{\sigma}$ correspond à l'estimateur de la variance des résidus pour le modèle comprenant toutes les variables explicatives. Plus le coefficient Cp est proche de $p*$ moins le modèle est biaisé.

- **AIC** : le critère d'information d'Akaike (Akaike's Information Criterion) est défini par

$$AIC = W \ln(\frac{SCE}{W}) + 2p^*$$

Ce critère proposé par Akaike (1973) dérive de la théorie de l'information, et s'appuie sur la mesure de Kullback et Leibler (1951). C'est un critère de sélection de modèles qui pénalise les modèles pour lesquels l'ajout de nouvelles variables explicatives n'apporte pas suffisamment d'information au modèle, l'information étant mesurée au travers de la SCE. On cherche à minimiser le critère AIC.

- **SBC** : le critère bayésien de Schwarz (Schwarz's Bayesian Criterion) est défini par

$$SBC = W \ln(\frac{SCE}{W}) + \ln(W)p^*$$

Ce critère proposé par Schwarz (1978) est proche du critère AIC, et comme ce dernier on cherche à le minimiser.

- **PC** : le critère de prédiction d'Amemiya (Amemiya's Prediction Criterion) est défini par

$$PC = \frac{(1-R^2)(W+p^*)}{W-p^*}$$

Ce critère proposé par Amemiya (1980) permet comme le R^2 ajusté de tenir compte de la parcimonie du modèle.

Le **tableau d'analyse de la variance** permet d'évaluer le pouvoir explicatif des variables explicatives. Le pouvoir explicatif est évalué en comparant l'ajustement (au sens des moindres carrés) du modèle final avec l'ajustement du modèle rudimentaire composé d'une constante égale à la moyenne de la variable dépendante.

Le tableau des **Type I SS** permet de visualiser l'influence de l'ajout progressif des variables explicatives sur l'ajustement du modèle, au sens de la somme des carrés des erreurs (SCE), de la moyenne des carrés des erreurs (MCE), du F de Fisher, ou de la probabilité associée au F de Fisher. Plus la probabilité est faible, plus la contribution de la variable au modèle est importante, toutes les autres variables étant déjà dans le modèle. Remarques : l'ordre de sélection des variables dans le modèle influe sur les valeurs obtenues ; la somme des sommes des carrés de ce tableau est égale à la somme des carrés du modèle.

Le tableau des **Type II SS** permet de visualiser l'influence du retrait d'une variable explicative sur l'ajustement du modèle, toutes les autres variables étant conservées, au sens de la somme des carrés des erreurs (SCE), de la moyenne des carrés des erreurs (MCE), du F de Fisher ou

de la probabilité associée au F de Fisher. Plus la probabilité est faible, plus la contribution de la variable au modèle est importante. Remarque : dans le cas des ANOVAs déséquilibrées, l'utilisation des Type III est recommandée mais XLSTAT affiche les Type II pour les utilisateurs avancés qui voudraient disposer des Type II.

Le tableau des **Type III SS** permet de visualiser l'influence du retrait d'une variable explicative sur l'ajustement du modèle, toutes les autres variables étant conservées, au sens de la somme des carrés des erreurs (SCE), de la moyenne des carrés des erreurs (MCE), du F de Fisher ou de la probabilité associée au F de Fisher. Plus la probabilité est faible, plus la contribution de la variable au modèle est importante, toutes les autres variables étant déjà dans le modèle. Remarque : contrairement au cas des Type I SS, l'ordre de sélection des variables dans le modèle n'influe pas sur les valeurs obtenues, et contrairement aux Type II SS, les valeurs ne dépendent pas des effectifs des cellules (par cellule on entend une combinaison de modalités des différents facteurs), ce qui fait des Type III le test recommandé pour évaluer la contribution d'une variable.

L'**équation du modèle** est ensuite affichée pour faciliter la lecture ou la réutilisation du modèle.

Le tableau des **coefficients normalisés** (aussi appelés coefficients bêta) permet de comparer le poids relatif des variables. Plus la valeur absolue d'un coefficient est élevée, plus le poids de la variable correspondante est important. Lorsque l'intervalle de confiance autour des coefficients normalisés comprend la valeur 0 (cela est facilement visible sur le graphique des coefficients normalisés), le poids d'une variable dans le modèle n'est pas significatif.

Exemple

Un exemple d'analyse d'un plan de Taguchi est disponible sur le Centre d'aide XLSTAT :

http://www.xlstat.com/demo-taguchif.htm

Bibliographie

Taguchi, G., & Wu, Y. (1980). Introduction to Off-Line Quality Control. Central Japan Quality Control Association.

Taguchi, G., Chowdhury, S., Wu, Y. & Yano, H. (2005). Taguchi's Quality Engineering Handbook. Hoboken, N.J., John Wiley & Sons.

Sabre, R. (2007). Plans d'expériences - Méthode de Taguchi. Techn. l'Ingénieur, Tech. Rep. F1006 V1.

Analyse de survie

Analyse de Kaplan-Meier

Utilisez cet outil pour créer des courbes de survie en utilisant la méthode de Kaplan-Meier (aussi appelée *product-limit*), et pour obtenir des informations essentielles comme le temps médian de survie. La méthode de Kaplan-Meier permet d'estimer les fonctions de survie, sans que les intervalles de temps soient nécessairement réguliers, contrairement à la méthode des tables actuarielles. XLSTAT permet le traitement de données censurées et de comparer différents groupes au sein de la population.

Dans cette section :

Description

La méthode de Kaplan-Meier permet d'obtenir rapidement une courbe de survie, ainsi que des statistiques essentielles comme le temps médian résiduel de survie. La méthode de Kaplan-Meier permet d'estimer les fonctions de survie, sans nécessiter que les intervalles de temps soient réguliers, contrairement à la méthode des tables actuarielles de survie.

Les courbes de survie permettent d'analyser l'évolution de l'effectif d'une population donnée avec le temps. Cette technique est utilisée pour l'analyse de données de survie, qu'il s'agisse d'individus (recherche sur le cancer par exemple), ou de produits (résistance au temps d'un outil de production par exemple) : certains individus meurent (les produits cassent), mais d'autres sortent de l'étude parce qu'ils guérissent, que l'on perd leur trace (déménagement par exemple) ou parce que l'étude est interrompue. Le premier type d'information est appelé « données événement », tandis que le second est appelé « données censurées ».

Il existe plusieurs types de censure pour les données de survie :

- Censure à gauche : lorsqu'un événement est enregistré au temps $t = t(i)$, cela signifie qu'il a eu lieu à $t < t(i)$.
- Censure à droite : lorsqu'un événement est enregistré au temps $t = t(i)$, cela signifie qu'il a eu lieu à $t > t(i)$, s'il n'a jamais eu lieu.

- Censure par intervalle : lorsqu'un événement est enregistré au temps $t = t(i)$, cela signifie qu'il a eu lieu pendant l'intervalle de temps $[t(i-1); t(i)]$.
- Censure exacte : lorsqu'un événement est enregistré au temps $t = t(i)$, cela signifie qu'il a eu lieu exactement à $t = t(i)$.

L'utilisation de la méthode Kaplan-Meier implique que l'on fait l'hypothèse que les observations sont indépendantes. De même, on fait l'hypothèse que la censure est indépendante : soient deux individus pris au hasard, inclus dans l'étude au temps $t-1$; si l'un deux est censuré au temps t, alors leur chance de survie est égale au temps t. On distingue quatre types de censure indépendante :

- Type I simple : tous les individus sont censurés après une même durée.
- Type I progressif : tous les individus sont censurés à la même date, quelle que soit la durée pendant laquelle ils ont été suivis (fin de l'étude par exemple).
- Type II : les individus sont suivis jusqu'à ce que l'on ait observé n événements.
- Aléatoire : le temps auquel se produit une censure est indépendant du temps de survie.

Si les « données événement » sont souvent mesurées par intervalle ou à une date exacte, les « données censurées » sont quant à elles, en général censurées à droite, la censure étant indépendante et aléatoire.

La méthode de Kaplan-Meier permet aussi de comparer des populations, en s'appuyant sur leur courbe de survie. Par exemple, il peut être intéressant de comparer les temps de survie des hommes et des femmes face à une même maladie, ou de comparer les temps de casse pour un même produit fabriqué sur deux chaînes de production différentes.

Intervalle de confiance

Le calcul des intervalles de confiance associés à la fonction de survie peut se faire de différentes manières.

Méthode de Greenwood :

$$S(T) \pm z_{1-\alpha/2}\sqrt{\frac{\mathrm{Var}(S(T))}{S^2(T)}}$$

Méthode exponentielle de Greenwood :

$$\exp(-\exp(\log(-\log(S(T))) \pm z_{1-\alpha/2}\sqrt{\mathrm{var}(S(T))}))$$

Méthode du log-transformé : $S(T)^{1/\theta}, S(T)^{\theta}$ avec :

$$\theta = exp\left(\frac{z_{1-\alpha/2}\sqrt{\frac{\mathrm{Var}(S(T))}{S^2(T)}}}{log(S(T))}\right)$$

Ces 3 méthodes donnent des résultats proches, mais on préférera les deux dernières dans le cas de petits échantillons.

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.

OK : cliquez sur ce bouton pour lancer les calculs.

Annuler : cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

Aide : cliquez sur ce bouton pour afficher l'aide.

 : cliquez sur ce bouton pour rétablir les options par défaut.

 : cliquez sur ce bouton pour effacer les sélections de données.

Onglet **Général**:

Données de dates : sélectionnez les données correspondant aux dates auxquelles se produisent les événements ou les censures. Si un en-tête a été sélectionné sur la première ligne, veillez à ce que l'option « libellés des colonnes » soit activée.

Données pondérées : activez cette option si, pour un temps donné, plusieurs événements ont pu être enregistrés (par exemple, au temps 218, 10 décès et 2 données censurées ont été enregistrés). Si vous activez cette option, « Indicateur d'événement » remplace « indicateur d'état », et l' « indicateur de censure » remplace les « Code événement » et « Code censuré ».

Indicateur d'état : sélectionnez ici les données correspondant à une « donnée événement » ou à une « donnée censurée ». Ce champ n'est pas disponible si l'option « Données pondérées » est activée. Si un en-tête a été sélectionné sur la première ligne, veillez à ce que l'option « libellés des colonnes » soit activée.

Code événement : entrez le code utilisé pour identifier une « donnée événement ». La valeur par défaut est 1.

Code censuré : entrez ici le code utilisé pour identifier une « donnée censurée ». La valeur par défaut est 0.

Indicateur d'événement : sélectionnez ici les données correspondant aux comptages des événements enregistrés à chaque temps. Ce champ n'est disponible que si l'option « Données pondérées » est activée. Si un en-tête a été sélectionné sur la première ligne, veillez à ce que l'option « libellés des colonnes » soit activée.

Indicateur de censure : sélectionnez ici les données correspondant aux comptages des données censurées enregistrées à chaque temps. Ce champ n'est disponible que si l'option « Données pondérées » est activée. Si un en-tête a été sélectionné sur la première ligne, veillez à ce que l'option « libellés des colonnes » soit activée.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés inclus : activez cette option si la première ligne et la première colonne des données sélectionnées contient un libellé.

Onglet **Options**:

Niveau de signification (%) : entrez la valeur du niveau de signification à utiliser pour les tests (valeur par défaut : 5%). Cette valeur est aussi utilisée pour déterminer les intervalles de confiance pour les statistiques calculées.

Intervalle de confiance : choisir le type d'intervalle de confiance qui sera calculé dans le tableau de sortie (voir la partie description pour des détails sur le calcul des intervalles de confiance).

Onglet **Prétraitement** :

Données manquantes :

- **Ne pas accepter les valeurs manquantes** : activez cette option pour que XLSTAT empêche la poursuite des calculs si des valeurs manquantes sont détectées.
- **Supprimer les observations** : activez cette option pour supprimer les observations comportant des données manquantes.

Groupes :

Analyse par groupe : activez cette option puis sélectionnez ici les données d'appartenance à des groupes si vous souhaitez que les calculs soient effectués sur chaque groupe séparément.

- **Comparer** : activez cette option si vous souhaitez que les courbes soient comparées pour les différents groupes, et si vous souhaitez que les tests de comparaison soient calculés.

Filtrer : activez cette option puis sélectionnez ici les données d'appartenance à des groupes, si vous souhaitez que les calculs ne soient effectués que sur certains groupes. Une boîte de dialogue apparaîtra au début des calculs, pour vous permettre de choisir les groupes actifs. Si l'option « Analyse par groupe » est activée, l'analyse ne sera effectuée groupe par groupe, que pour les groupes sélectionnés.

Onglet **Graphiques** :

Fonction de survie cumulée : activez cette option pour afficher les graphiques relatifs à la fonction de survie cumulée.

-Log(FSC) : activez cette option si vous souhaitez que XLSTAT affiche le –Log() de la fonction de survie (FSC).

Log(-Log(FSC)) : activez cette option si vous souhaitez que XLSTAT affiche le Log(–Log()) de la fonction de survie (FSC).

1 - Fonction de survie cumulée : activez cette option si vous souhaitez que XLSTAT affiche le graphique représentant 1 - la fonction de survie cumulée (FSC).

Données censurées : activez cette option si vous souhaitez que les données pour lesquelles des données censurées ont été observées soient identifiées sur le graphique (vous avez le choix entre un « o » et un « + » pour l'identification).

Résultats

Statistiques simples : vous trouverez dans ce tableau le nombre total d'individus pris en compte dans l'analyse, le nombre d'événements, et le nombre de données censurées.

Tableau de Kaplan-Meier : dans ce tableau sont affichés plusieurs résultats :

- **Début de l'intervalle** : borne inférieure de l'intervalle de temps.
- **A risque** : nombre d'individus à risque.
- **Événements** : nombre d'événements enregistrés.
- **Censurées** : nombre de données censurées enregistrées.
- **Proportion d'événements** : proportion d'individus qui n'a pas survécu.
- **Taux de survie** : proportion d'individus qui a survécu.
- **Fonction de survie (FSC)** : probabilité pour un individu de survivre au moins jusqu'au temps considéré.
- **Écart-type de la fonction de survie** : écart-type de la quantité précédente.
- **Intervalle de confiance de la fonction de survie** : intervalle de confiance de la quantité précédente.

Temps moyen et médian de survie : dans le premier tableau sont affichés le temps moyen résiduel de survie et l'écart-type correspondant. Dans un second tableau sont affichés le temps résiduel pour trois quartiles au début de l'expérience. La médiane correspond au quartile 50%. Un intervalle de confiance sur ces statistiques est aussi fourni.

Graphiques : en fonction des options choisies, jusqu'à trois graphiques peuvent être affichés : Fonction de survie cumulée (FSC), -Log(FSC), Log(-Log(FSC)) et 1- FSC.

Si l'option "Comparer" a été activée, XLSTAT affiche les résultats suivants :

Tests d'égalité des fonctions de survie : ce tableau affiche les statistiques correspondant à trois tests : le Log-rank test, le test de Wilcoxon, et le test de Tarone Ware. Ces tests utilisent la distribution du Khi². Plus la p-value est faible, plus la différence entre les courbes est significative.

Si la pvalue obtenue par le test du log-rank est significative au seuil α = 5%, des tests de comparaisons multiples sont effectués sur les groupes 2 à 2. Nous utilisons dans ce cas le test de Dunn-Sidak qui est un dérivé du test de Bonferroni et est plus performant dans certaines situations.

$$\alpha' = 1 - (1 - \alpha)^{1/g}.$$

où g est le nombre de groupes comparés.

Graphiques : en fonction des options choisies, jusqu'à 4 graphiques peuvent être affichés, avec pour chacun, une courbe par groupe : Fonction de survie (FSC), -Log(FSC), Log(-Log(FSC)) et 1- FSC.

Exemple

Un exemple d'analyse de survie par la méthode de Kaplan-Meier est disponible sur le Centre d'aide XLSTAT :

http://www.xlstat.com/demo-kmf.htm

Bibliographie

Brookmeyer R. and Crowley J. (1982). A confidence interval for the median survival time. *Biometrics*, **38**, 29-41.

Collett D. (1994). Modeling Survival Data In Medical Research. Chapman and Hall, London.

Cox D.R. and Oakes D. (1984). Analysis of Survival Data. Chapman and Hall, London.

Elandt-Johnson R.C. and Johnson N.L. (1980). Survival Models and Data Analysis. John Wiley & Sons, New York.

Kalbfleisch J.D. and Prentice R.L. (1980). The Statistical Analysis of Failure Time Data. John Wiley & Sons, New York.

Tableaux de survie

Utilisez cet outil pour créer des courbes de survie, et pour obtenir des informations essentielles comme le temps médian de survie. L'analyse des tableaux actuariels, se fonde sur des intervalles de temps réguliers, contrairement à la méthode de Kaplan-Meier qui traite les événements au fur et à mesure de leur apparition. XLSTAT permet le traitement de données censurées et la comparaison de différents groupes au sein de la population.

Dans cette section :

Description

L'analyse de tableaux actuariels appartient aux méthodes descriptives de l'analyse de survie, de même que l'analyse de Kaplan-Meier, méthode plus récente et qui s'avère plus performante dans certaines conditions.

L'analyse de tableaux actuariels permet d'obtenir rapidement une courbe de survie, ainsi que des statistiques essentielles comme le temps médian résiduel de survie.

Les tables actuarielles permettent d'analyser l'évolution de l'effectif d'une population donnée avec le temps. Cette technique est utilisée pour l'analyse de données de survie, qu'il s'agisse d'individus (recherche sur le cancer par exemple), ou de produits (résistance au temps d'un outil de production par exemple) : certains individus meurent (les produits cassent), mais d'autres sortent de l'étude parce qu'ils guérissent, que l'on perd leur trace (déménagement par exemple) ou parce que l'étude est interrompue. Le premier type d'information est appelé « données événement », tandis que le second est appelé « données censurées ».

Il existe plusieurs types de censure pour les données de survie :

Censure à gauche : lorsqu'un événement est enregistré au temps t=t(i), cela signifie qu'il a eu lieu à t < t(i).

Censure à droite : lorsqu'un événement est enregistré au temps t=t(i), cela signifie qu'il a eu lieu à t > t(i), s'il n'a jamais eu lieu.

Censure par intervalle : lorsqu'un événement est enregistré au temps t=t(i), cela signifie qu'il a eu lieu pendant l'intervalle de temps [t(i-1); t(i)].

Censure exacte : lorsqu'un événement est enregistré au temps t=t(i), cela signifie qu'il a eu lieu exactement à t=t(i).

La méthode des tables actuarielles implique l'hypothèse que les observations sont indépendantes. De même, on fait l'hypothèse que la censure est indépendante : soient deux individus pris au hasard, inclus dans l'étude au temps t-1 ; si l'un deux est censuré au temps t, alors leur chance de survie est égale au temps t. On distingue quatre types de censure indépendante :

Type I simple : tous les individus sont censurés après une même durée.

Type I progressif : tous les individus sont censurés à la même date, quelle que soit la durée pendant laquelle ils ont été suivis (fin de l'étude par exemple).

Type II : les individus sont suivis jusqu'à ce que l'on ait observé n événements.

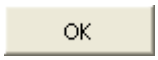
Aléatoire : le temps auquel se produit une censure est indépendant du temps de survie.

Si les « données événement » sont souvent mesurées par intervalle ou à une date exacte, les « données censurées » sont quant à elles, en général censurées à droite, la censure étant indépendante et aléatoire.

L'analyse de tableaux actuariels permet de comparer des populations, en s'appuyant sur leur courbe de survie. Par exemple, il peut être intéressant de comparer les temps de survie des hommes et des femmes face à une même maladie, ou de comparer les temps de casse pour un même produit fabriqué sur deux chaînes de production différentes.

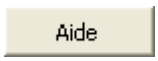
Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.



: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.

Onglet **Général**:

Données de dates : sélectionnez les données correspondant aux dates auxquelles se produisent les événements ou les censures. Si un en-tête a été sélectionné sur la première ligne, veillez à ce que l'option « libellés des colonnes » soit activée.

Données pondérées : activez cette option si, pour un temps donné, plusieurs événements ont pu être enregistrés (par exemple, au temps 218, 10 décès et 2 données censurées ont été enregistrés). Si vous activez cette option, « Indicateur d'événement » remplace « indicateur d'état », et l' « indicateur de censure » remplace les « Code événement » et « Code censuré ».

Indicateur d'état : sélectionnez ici les données correspondant à une « donnée événement » ou à une « donnée censurée ». Ce champ n'est pas disponible si l'option « Données pondérées » est activée. Si un en-tête a été sélectionné sur la première ligne, veillez à ce que l'option « libellés des colonnes » soit activée.

Code événement : entrez le code utilisé pour identifier une « donnée événement ». La valeur par défaut est 1.

Code censuré : entrez ici le code utilisé pour identifier une « donnée censurée ». La valeur par défaut est 0.

Indicateur d'événement : sélectionnez ici les données correspondant aux comptages des événements enregistrés à chaque temps. Ce champ n'est disponible que si l'option « Données pondérées » est activée. Si un en-tête a été sélectionné sur la première ligne, veillez à ce que l'option « libellés des colonnes » soit activée.

Indicateur de censure : sélectionnez ici les données correspondant aux comptages des données censurées enregistrées à chaque temps. Ce champ n'est disponible que si l'option « Données pondérées » est activée. Si un en-tête a été sélectionné sur la première ligne, veillez à ce que l'option « libellés des colonnes » soit activée.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés inclus : activez cette option si la première ligne et la première colonne des données sélectionnées contient un libellé.

Onglet **Options**:

Niveau de signification (%) : entrez la valeur du niveau de signification à utiliser pour les tests (valeur par défaut : 5%). Cette valeur est aussi utilisée pour déterminer les intervalles de confiance pour les statistiques calculées.

Intervalles de temps :

- **Amplitude constante** : activez cette option, puis entrez l'amplitude des intervalles de temps à utiliser pour l'analyse. La valeur par défaut est 1.
- **Définis par l'utilisateur** : activez cette option, puis sélectionnez la borne inférieure du premier intervalle de temps et les bornes supérieures de tous les intervalles de temps que vous voulez utiliser pour l'analyse.

Onglet **Prétraitement** :

Données manquantes :

Ne pas accepter les valeurs manquantes : activez cette option pour que XLSTAT empêche la poursuite des calculs si des valeurs manquantes sont détectées.

Supprimer les observations : activez cette option pour supprimer les observations comportant des données manquantes.

Groupes :

Analyse par groupe : activez cette option puis sélectionnez ici les données d'appartenance à des groupes si vous souhaitez que les calculs soient effectués sur chaque groupe séparément.

- **Comparer** : activez cette option si vous souhaitez que les courbes soient comparées pour les différents groupes, et si vous souhaitez que les tests de comparaison soient calculés.

Filtrer : activez cette option puis sélectionnez ici les données d'appartenance à des groupes, si vous souhaitez que les calculs ne soient effectués que sur certains groupes. Une boîte de dialogue apparaîtra au début des calculs, pour vous permettre de choisir les groupes actifs. Si l'option « Analyse par groupe » est activée, l'analyse ne sera effectuée groupe par groupe, que pour les groupes sélectionnés.

Onglet **Graphiques** :

Fonction de survie cumulée : activez cette option pour afficher les graphiques relatifs à la fonction de survie cumulée.

-Log(FSC) : activez cette option si vous souhaitez que XLSTAT affiche le –Log() de la fonction de survie (FSC).

Log(-Log(FSC)) : activez cette option si vous souhaitez que XLSTAT affiche le Log(–Log()) de la fonction de survie (FSC).

Données censurées : activez cette option si vous souhaitez que les données pour lesquelles des données censurées ont été observées soient identifiées sur le graphique (vous avez le choix entre un « o » et un « + » pour l'identification)

Résultats

Statistiques simples : vous trouverez dans ce tableau le nombre total d'individus pris en compte dans l'analyse, le nombre d'événements, et le nombre de données censurées.

Table de survie : dans ce tableau sont affichés les résultats suivants :

- **Intervalle** : intervalle de temps.
- **A risque** : nombre d'individus à risque pendant l'intervalle de temps.
- **Evénements** : nombre d'événements enregistrés pendant l'intervalle de temps.
- **Censurées** : nombre de données censurées enregistrées pendant l'intervalle de temps.
- **Effectivement à risque** : nombre d'individus effectivement à risque pendant l'intervalle de temps.
- **Taux de survie** : proportion d'individus qui ont survécu (l'événement ne s'est pas produit) pendant l'intervalle de temps. Ratio des individus qui ont survécu sur les individus effectivement à risque.
- **Probabilité conditionnelle d'événement** : ratio des individus qui ont n'ont pas survécu sur les individus effectivement à risque.
- Ecart-type de la probabilité conditionnelle d'événement : écart-type de la quantité précédente.
- **Fonction de survie (FSC)** : probabilité pour un individu de survivre au moins jusqu'au temps considéré.
- **Ecart-type de la fonction de survie** : écart-type de la quantité précédente.
- **Densité de probabilité** : fonction de densité estimée au milieu de l'intervalle de temps considéré.
- **Ecart-type de la densité de probabilité** : écart-type de la quantité précédente.
- **Taux de hasard** : estimation du taux de hasard au milieu de l'intervalle de temps considéré. Cet indicateur, aussi appelé taux d'échec, correspond au taux d'échec observé pour les survivants.
- **Ecart-type du taux de hasard** : écart-type de la quantité précédente.
- **Temps médian résiduel de survie** : quantité de temps restant pour réduire la taille de la population de 50% (individus à risque).
- Ecart-type du temps médian résiduel de survie : écart-type de la quantité précédente.

Temps de survie médian : vous trouverez dans ce tableau le temps médian résiduel de survie au début de l'expérience, ainsi que l'écart-type de ce dernier. Cette statistique permet d'évaluer le temps au bout duquel la taille de la population étudiée a réduit de moitié.

Graphiques : en fonction des options choisies, jusqu'à cinq graphiques peuvent être affichés : Fonction de survie cumulative (FSC), densité de probabilité, Taux de hasard, -Log(FSC) et Log(-Log(FSC)).

Si l'option "Comparer" a été activée, XLSTAT affiche les résultats suivants :

Tests d'égalité des fonctions de survie : ce tableau affiche les statistiques correspondant à trois tests : le Log-rank test, le test de Wilcoxon, et le test de Tarone Ware test. Ces tests s'appuient tous sur le test du Khi². Plus la p-value est faible, plus la différence entre les courbes est significative.

Si la pvalue obtenue par le test du log-rank est significative au seuil alpha = 5%, des tests de comparaisons multiples sont effectués sur les groupes 2 à 2. Nous utilisons dans ce cas le test de Dunn-Sidak qui est un dérivé du test de Bonferroni et est plus performant dans certaines situations.

$$\alpha' = 1 - (1 - \alpha)^{1/g}.$$

où g est le nombre de groupes comparés.

Graphiques : en fonction des options choisies, jusqu'à cinq graphiques peuvent être affichés, avec pour chacun une courbe par groupe : Fonction de survie (FSC), densité de probabilité, Taux de hasard, -Log(FSC) et Log(-Log(FSC)).

Exemple

Un exemple d'analyse de survie par la méthode des tables actuarielles est disponible sur le Centre d'aide XLSTAT :

http://www.xlstat.com/demo-lifef.htm

Bibliographie

Brookmeyer R. and Crowley J. (1982). A confidence interval for the median survival time. *Biometrics*, **38**, 29-41.

Collett D. (1994). Modeling Survival Data In Medical Research. Chapman and Hall, London.

Cox D.R. and Oakes D. (1984). Analysis of Survival Data. Chapman and Hall, London.

Elandt-Johnson R.C. and Johnson N.L. (1980). Survival Models and Data Analysis. John Wiley & Sons, New York.

Kalbfleisch J.D. and Prentice R.L. (1980). The Statistical Analysis of Failure Time Data. John Wiley & Sons, New York.

Analyse de Nelson-Aalen

Utilisez cet outil pour créer des courbes de risque en utilisant la méthode de Nelson-Aalen. La méthode de Nelson-Aalen permet d'estimer les fonctions de risque, sans que les intervalles de temps soient nécessairement réguliers, contrairement à la méthode des tables actuarielles. XLSTAT permet le traitement de données censurées et de comparer différents groupes au sein de la population.

Dans cette section :

Description

La méthode de Nelson-Aalen permet d'obtenir rapidement une courbe de risque cumulé. La méthode de Nelson-Aalen permet d'estimer les fonctions de risque, sans que les intervalles de temps soient nécessairement réguliers, contrairement à la méthode des tables actuarielles.

Les courbes de risque cumulé permettent d'analyser l'évolution du risque avec le temps. Cette technique est utilisée pour l'analyse de données de survie, qu'il s'agisse d'individus (recherche sur le cancer par exemple), ou de produits (résistance au temps d'un outil de production par exemple) : certains individus meurent (les produits cassent), mais d'autres sortent de l'étude parce qu'ils guérissent, que l'on perd leur trace (déménagement par exemple) ou parce que l'étude est interrompue. Le premier type d'information est appelé « données événement », tandis que le second est appelé « données censurées ».

Il existe plusieurs types de censure pour les données de survie :

- **Censure à gauche** : lorsqu'un événement est enregistré au temps $t = t(i)$, cela signifie qu'il a eu lieu à $t < t(i)$.
- **Censure à droite** : lorsqu'un événement est enregistré au temps $t = t(i)$, cela signifie qu'il a eu lieu à $t > t(i)$, s'il n'a jamais eu lieu.
- **Censure par intervalle** : lorsqu'un événement est enregistré au temps $t = t(i)$, cela signifie qu'il a eu lieu pendant l'intervalle de temps $[t(i-1); t(i)]$.
- **Censure exacte** : lorsqu'un événement est enregistré au temps $t = t(i)$, cela signifie qu'il a eu lieu exactement à $t = t(i)$.

L'utilisation de la méthode Nelson-Aalen implique que l'on fait l'hypothèse que les observations sont indépendantes. De même, on fait l'hypothèse que la censure est indépendante : soient deux individus pris au hasard, inclus dans l'étude au temps $t-1$; si l'un d'eux est censuré au temps t, alors leur chance de survie est égale au temps t. On distingue quatre types de censure indépendante :

- **Type I simple** : tous les individus sont censurés après une même durée.
- **Type I progressif** : tous les individus sont censurés à la même date, quelle que soit la durée pendant laquelle ils ont été suivis (fin de l'étude par exemple).
- **Type II** : les individus sont suivis jusqu'à ce que l'on ait observé n événements.
- **Aléatoire** : le temps auquel se produit une censure est indépendant du temps de survie.

Si les « données événement » sont souvent mesurées par intervalle ou à une date exacte, les « données censurées » sont quant à elles, en général censurées à droite, la censure étant indépendante et aléatoire.

La méthode de Nelson-Aalen permet aussi de comparer le risque associé à des populations, en s'appuyant sur leur courbe de risque.

La méthode de Nelson-Aalen est la plus adaptée pour calculer le risque cumulé au temps T, noté généralement $H(T)$. Si on s'intéresse au risque cumulé, elle est plus adaptée que la méthode de Kaplan-Meier. Si par contre on s'intéresse à la fonction de survie, on préfèrera la méthode de Kaplan-Meier à celle de Nelson-Aalen.

La fonction de risque cumulée est donnée par : $H(T) = \sum_{T_i \leq T} \frac{d_i}{r_i}$

Avec d_i, nombre d'évènement au temps T_i et r_i nombre d'individus toujours dans l'étude (à risques).

Plusieurs variances existent pour $H(T)$:

- Variance simple : $\mathrm{var}(H(T)) = \sum_{T_i \leq T} \frac{d_i}{r_i^2}$
- Variance plug-in : $\mathrm{var}(H(T)) = \sum_{T_i \leq T} \frac{d_i\,(r_i - d_i)}{r_i^3}$
- Variance binomiale : $\mathrm{var}(H(T)) = \sum_{T_i \leq T} \frac{d_i\,(r_i - d_i)}{r_i^2\,(r_i - 1)}$

Des intervalles de confiance peuvent être obtenus :

- Méthode de Greenwood : $H(T) \pm z_{1-\alpha/2}\sqrt{\mathrm{Var}(H(T))}$

- Méthode du log-transformé : $H(T)/\phi,\ H(T) \times \phi$ avec $\phi = \exp(\frac{z_{1-\alpha/2}\sqrt{\text{var}(H(T))}}{H(T)})$

Ces 2 méthodes donnent des résultats proches, mais on préférera la seconde dans le cas de petits échantillons.

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.

OK : cliquez sur ce bouton pour lancer les calculs.

Annuler : cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

Aide : cliquez sur ce bouton pour afficher l'aide.

 : cliquez sur ce bouton pour rétablir les options par défaut.

 : cliquez sur ce bouton pour effacer les sélections de données.

Onglet **Général**:

Données de dates : sélectionnez les données correspondant aux dates auxquelles se produisent les événements ou les censures. Si un en-tête a été sélectionné sur la première ligne, veillez à ce que l'option « libellés des colonnes » soit activée.

Données pondérées : activez cette option si, pour un temps donné, plusieurs événements ont pu être enregistrés (par exemple, au temps 218, 10 décès et 2 données censurées ont été enregistrés). Si vous activez cette option, « Indicateur d'événement » remplace « indicateur d'état », et l' « indicateur de censure » remplace les « Code événement » et « Code censuré ».

Indicateur d'état : sélectionnez ici les données correspondant à une « donnée événement » ou à une « donnée censurée ». Ce champ n'est pas disponible si l'option « Données pondérées » est activée. Si un en-tête a été sélectionné sur la première ligne, veillez à ce que l'option « libellés des colonnes » soit activée.

Code événement : entrez le code utilisé pour identifier une « donnée événement ». La valeur par défaut est 1.

Code censuré : entrez ici le code utilisé pour identifier une « donnée censurée ». La valeur par défaut est 0.

Indicateur d'événement : sélectionnez ici les données correspondant aux comptages des événements enregistrés à chaque temps. Ce champ n'est disponible que si l'option « Données pondérées » est activée. Si un en-tête a été sélectionné sur la première ligne, veillez à ce que l'option « libellés des colonnes » soit activée.

Indicateur de censure : sélectionnez ici les données correspondant aux comptages des données censurées enregistrées à chaque temps. Ce champ n'est disponible que si l'option « Données pondérées » est activée. Si un en-tête a été sélectionné sur la première ligne, veillez à ce que l'option « libellés des colonnes » soit activée.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés inclus : activez cette option si la première ligne des données sélectionnées contient un libellé.

Onglet **Options**:

Niveau de signification (%) : entrez la valeur du niveau de signification à utiliser pour les tests (valeur par défaut : 5%). Cette valeur est aussi utilisée pour déterminer les intervalles de confiance pour les statistiques calculées.

Variance : choisir le type de variance qui sera calculée dans le tableau de sortie (voir la partie description pour des détails sur le calcul des variances).

Intervalle de confiance : choisir le type d'intervalle de confiance qui sera calculé dans le tableau de sortie (voir la partie description pour des détails sur le calcul des intervalles de confiance).

Onglet **Prétraitement** :

Données manquantes :

Ne pas accepter les valeurs manquantes : activez cette option pour que XLSTAT empêche la poursuite des calculs si des valeurs manquantes sont détectées.

Supprimer les observations : activez cette option pour supprimer les observations comportant des données manquantes.

Groupes :

Analyse par groupe : activez cette option puis sélectionnez ici les données d'appartenance à des groupes si vous souhaitez que les calculs soient effectués sur chaque groupe séparément.

- **Comparer** : activez cette option si vous souhaitez que les courbes soient comparées pour les différents groupes, et si vous souhaitez que les tests de comparaison soient calculés.

Filtrer : activez cette option puis sélectionnez ici les données d'appartenance à des groupes si vous souhaitez que les calculs ne soient effectués que sur certains groupes. Une boîte de dialogue sera affichée pour vous permettre de choisir les groupes actifs. Si l'option « Analyse par groupe » est activée, l'analyse ne sera effectuée groupe par groupe que pour les groupes sélectionnés.

Onglet **Graphiques** :

Fonction de risque cumulée : activez cette option pour afficher les graphiques relatifs à la fonction de risque cumulée.

Fonction de survie cumulée : activez cette option si vous souhaitez que XLSTAT affiche la fonction de survie.

Log(Fonction de risque) : activez cette option si vous souhaitez que XLSTAT affiche le Log() de la fonction de risque.

Données censurées : activez cette option si vous souhaitez que les données pour lesquelles des données censurées ont été observées soient identifiées sur le graphique (vous avez le choix entre un « o » et un « + » pour l'identification)

Résultats

Statistiques simples : vous trouverez dans ce tableau le nombre total d'individus pris en compte dans l'analyse, le nombre d'événements, et le nombre de données censurées.

Tableau de Nelson-Aalen : dans ce tableau sont affichés plusieurs résultats :

- **Début de l'intervalle** : borne inférieure de l'intervalle de temps.
- **A risque** : nombre d'individus à risque.
- **Evénements** : nombre d'événements enregistrés.
- **Censurées** : nombre de données censurées enregistrées.
- **Fonction de risque cumulé** : risque associé à un individu au temps considéré.
- **Ecart-type** : écart-type de la quantité précédente.
- **Intervalle de confiance** : intervalle de confiance de la quantité précédente.
- **Fonction de survie (FSC)** : probabilité pour un individu de survivre au moins jusqu'au temps considéré (calculée comme $S(T) = \exp(-H(T))$).

Graphiques : en fonction des options choisies, jusqu'à trois graphiques peuvent être affichés : Fonction de risque cumulée, Fonction de survie cumulée et Log(fonction de risque cumulée).

Si l'option "Comparer" a été activée, XLSTAT affiche les résultats suivants:

Tests d'égalité des fonctions de survie : ce tableau affiche les statistiques correspondant à trois tests : le Log-rank test, le test de Wilcoxon, et le test de Tarone Ware. Ces tests utilisent la distribution du Khi^2. Plus la p-value est faible, plus la différence entre les courbes est significative.

Si la pvalue obtenue par le test du log-rank est significative au seuil alpha = 5%, des tests de comparaisons multiples sont effectués sur les groupes 2 à 2. Nous utilisons dans ce cas le test de Dunn-Sidak qui est un dérivé du test de Bonferroni et est plus performant dans certaines situations.

$$\alpha' = 1 - (1 - \alpha)^{1/g}.$$

où g est le nombre de groupes comparés.

Graphiques : en fonction des options choisies, jusqu'à trois graphiques peuvent être affichés, avec pour chacun, une courbe par groupe : Fonction de risque cumulée, Fonction de survie et Log(Fonction de risque cumulée).

Exemple

Un exemple d'analyse de survie par la méthode de Nelson-Aalen est disponible sur le Centre d'aide XLSTAT :

http://www.xlstat.com/demo-naf.htm

Bibliographie

Collett D. (1994). Modeling Survival Data In Medical Research. Chapman and Hall, London.

Cox D.R. and Oakes D. (1984). Analysis of Survival Data. Chapman and Hall, London.

Elandt-Johnson R.C. and Johnson N.L. (1980). Survival Models and Data Analysis. John Wiley & Sons, New York.

Kalbfleisch J.D. and Prentice R.L. (1980). The Statistical Analysis of Failure Time Data. John Wiley & Sons, New York.

Incidence cumulée

Utilisez cet outil pour analyser des données de survie lorsque plusieurs types d'évènement sont possibles.

L'incidence cumulée permet d'estimer des incidences lorsque plusieurs évènements compétitifs peuvent survenir. Les intervalles de temps ne doivent pas être nécessairement réguliers. XLSTAT permet le traitement de données censurées à risques compétitifs et de comparer différents groupes au sein de la population.

Dans cette section :

Description

L'incidence cumulée permet d'estimer des incidences lorsque plusieurs évènements compétitifs peuvent survenir. Les intervalles de temps ne doivent pas être nécessairement réguliers. XLSTAT permet le traitement de données censurées à risques compétitifs et de comparer différents groupes au sein de la population.

Pour une période donnée, l'incidence cumulée est la probabilité qu'une observation toujours incluse dans l'analyse au début de la période et présente tout au long de celle-ci soit affectée par un évènement au cours de la période. Elle est surtout adaptée dans le cas de risques compétitifs, c'est-à- dire lorsque plusieurs types d'évènement peuvent survenir.

Cette technique est utilisée pour l'analyse de données de survie, qu'il s'agisse d'individus (recherche sur le cancer par exemple), ou de produits (résistance au temps d'un outil de production par exemple) : certains individus meurent (dans ce cas on pourra avoir 2 risques, celui de mourir de la maladie et d'autres causes de décès), les produits cassent (dans ce cas on pourra modéliser différent type de casse), mais d'autres sortent de l'étude parce qu'ils guérissent, que l'on perd leur trace (déménagement par exemple) ou parce que l'étude est interrompue. Le premier type d'information est appelé « données événement », tandis que le second est appelé « données censurées ».

Il existe plusieurs types de censure pour les données de survie :

- Censure à gauche : lorsqu'un événement est enregistré au temps $t = t(i)$, cela signifie qu'il a eu lieu à $t \times t(i)$.

- Censure à droite : lorsqu'un événement est enregistré au temps $t = t(i)$, cela signifie qu'il a eu lieu à $t \times t(i)$, s'il n'a jamais eu lieu.
- Censure par intervalle : lorsqu'un événement est enregistré au temps $t = t(i)$, cela signifie qu'il a eu lieu pendant l'intervalle de temps $[t(i-1); t(i)]$.
- Censure exacte : lorsqu'un événement est enregistré au temps $t = t(i)$, cela signifie qu'il a eu lieu exactement à $t = t(i)$.

L'incidence cumulée implique que l'on fait l'hypothèse que les observations sont indépendantes. De même, on fait l'hypothèse que la censure est indépendante : soient deux individus pris au hasard, inclus dans l'étude au temps $t-1$; si l'un deux est censuré au temps t, alors leur chance de survie est égale au temps t. On distingue quatre types de censure indépendante :

- Type I simple : tous les individus sont censurés après une même durée.
- Type I progressif : tous les individus sont censurés à la même date, quelle que soit la durée pendant laquelle ils ont été suivis (fin de l'étude par exemple).
- Type II : les individus sont suivis jusqu'à ce que l'on ait observé n événements.
- Aléatoire : le temps auquel se produit une censure est indépendant du temps de survie.

Si les « données événement » sont souvent mesurées par intervalle ou à une date exacte, les « données censurées » sont quant à elles, en général censurées à droite, la censure étant indépendante et aléatoire.

Les différents types d'évènements ne peuvent arriver qu'une seule fois, une fois que l'évènement a eu lieu, l'observation est sortie de l'analyse. On peut ainsi calculer le risque d'apparition d'un évènement en présence d'évènements compétitifs. XLSTAT permet de comparer les types d'évènements mais aussi de prendre en compte des groupes d'observations (en fonction du traitement administré, par exemple).

Calcul de l'incidence cumulée : $I_k(T) = \sum_{T_j \leq T} \hat{S}(T_{j-1}) \frac{d_{kj}}{n_j}$ pour l'évènement k au temps T. Avec $\hat{S}(T_{j-1})$ la fonction de survie obtenue par l'estimateur de Kaplan-Meier au temps $T-1$, d_{kj} est le nombre d'évènement du type k au temps T_j et n_j le nombre d'observations à risque (encore dans l'analyse) au temps T_j. La variance est donnée par :

$$
\begin{aligned}
Var(I_k(T)) = & \sum_{T_j \leq T} \left[\left(I_k(T) - I_k(T_j)\right)^2 \frac{d_j}{n_j(n_j - d_j)} \right] \\
& + \sum_{T_j \leq T} \left[\left(\hat{S}(T_{j-1})\right)^2 \frac{(n_j - d_j)}{n_j} \frac{d_{kj}}{n_j^2} \right] \\
& - 2 \sum_{T_j \leq T} \left[\left(I_k(T) - I_k(T_j)\right) \hat{S}(T_{j-1}) \frac{d_j}{n_j^2} \right]
\end{aligned}
$$

On calcule des intervalles de confiance en utilisant la formule suivante :

$$
I_k(T)^{exp\left(\frac{\pm z_{\alpha/2}\sqrt{Var(I_k(T))}}{I_k(T) log(I_k(T))}\right)}
$$

Test de Gray pour la comparaison de groupes

Le test de Gray est utilisé pour comparer des groupes d'observations dans le cadre du calcul de l'incidence cumulée. Lorsqu'on a des risques compétitifs, les tests de comparaison de l'analyse de survie ne peuvent pas être appliqués. Gray (1988) a développé un test pour ce cas spécifique. Ce test est basé sur un test de comparaison de k échantillons global associé à chaque état. Pour plus de détails sur ce test, on peut voir Gray (1988).

Une p-valeur pour chaque état est obtenue pour tous les groups étudiés.

Résultats

Statistiques simples : vous trouverez dans ce tableau le nombre total d'individus pris en compte dans l'analyse, le nombre d'événements, et le nombre de données censurées.

Chaque tableau apparaît pour chaque type d'évènement.

Incidence cumulée : dans ce tableau sont affichés plusieurs résultats :

- **Début de l'intervalle** : borne inférieure de l'intervalle de temps.
- **A risque** : nombre d'individus à risque.
- **Evénements i** : nombre d'événements du type i enregistrés.
- **Tous les événements** : nombre d'événements de tous les types enregistrés.
- **Censurées** : nombre de données censurées enregistrées.
- **Incidence cumulée** : Incidence cumulée pour l'évènement i.
- **Ecart-type** : écart-type de la quantité précédente.
- **Intervalle de confiance** : intervalle de confiance de la quantité précédente.

Fonction de survie cumulée : dans ce tableau sont affichés plusieurs résultats :

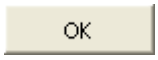
- **Début de l'intervalle** : borne inférieure de l'intervalle de temps.
- **A risque** : nombre d'individus à risque.
- **Evénements i** : nombre d'événements du type i enregistrés.
- **Tous les événements** : nombre d'événements de tous les types enregistrés.
- **Censurées** : nombre de données censurées enregistrées.
- **Fonction de survie cumulée** : Fonction de survie cumulée pour l'évènement i.
- **Ecart-type** : écart-type de la quantité précédente.
- **Intervalle de confiance** : intervalle de confiance de la quantité précédente.

Graphiques : en fonction des options choisies, jusqu'à deux graphiques peuvent être affichés : Incidence cumulée et fonction de survie cumulée.

Test de Gray : Pour chaque état, la statistique du test de Gray ainsi que le nombre de degré de liberté ainsi qu'une p-valeur sont affichés.

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.

Onglet **Général**:

Données de dates : sélectionnez les données correspondant aux dates auxquelles se produisent les événements ou les censures. Si un en-tête a été sélectionné sur la première ligne, veillez à ce que l'option « libellés des colonnes » soit activée.

Indicateur d'état : sélectionnez ici les données correspondant à une « donnée événement » ou à une « donnée censurée ». Si un en-tête a été sélectionné sur la première ligne, veillez à ce que l'option « libellés des colonnes » soit activée.

Code censuré : entrez ici le code utilisé pour identifier une « donnée censurée ». La valeur par défaut est 0.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés inclus : activez cette option si la première ligne des données sélectionnées contient un libellé.

Groupes : activez cette option puis sélectionnez ici les données d'appartenance à des groupes si vous souhaitez que les calculs soient effectués sur chaque groupe séparément.

Test de Gray : activez cette option si vous voulez comparer des courbes d'incidences cumulées à l'aide du test de Gray. Un test est alors appliqué pour chaque état.

Onglet **Options**:

Niveau de signification (%) : entrez la valeur du niveau de signification à utiliser pour les tests (valeur par défaut : 5%). Cette valeur est aussi utilisée pour déterminer les intervalles de confiance pour les statistiques calculées.

Onglet **Données manquantes** :

Ne pas accepter les valeurs manquantes : activez cette option pour que XLSTAT empêche la poursuite des calculs si des valeurs manquantes sont détectées.

Supprimer les observations : activez cette option pour supprimer les observations comportant des données manquantes.

Onglet **Graphiques** :

Incidence cumulée : activez cette option pour afficher les graphiques relatifs à l'incidence cumulée.

Fonction de survie cumulée : activez cette option si vous souhaitez que XLSTAT affiche la fonction de survie.

Données censurées : activez cette option si vous souhaitez que les données pour lesquelles des données censurées ont été observées soient identifiées sur le graphique (un « o » est utilisé pour l'identification).

Exemple

Un exemple d'analyse grâce à l'incidence cumulée est disponible sur le Centre d'aide XLSTAT :

http://www.xlstat.com/demo-cuif.htm

Bibliographie

Collett D. (1994). Modeling Survival Data In Medical Research. Chapman and Hall, London.

Cox D.R. and Oakes D. (1984). Analysis of Survival Data. Chapman and Hall, London.

Elandt-Johnson R.C. and Johnson N.L. (1980). Survival Models and Data Analysis. John Wiley & Sons, New York.

Gray, R.J. (1988) A class of K-sample tests for comparing the cumulative incidence of a competing risk, *The Annals of statistics*, **16(3)**, 1141-1154.

Kalbfleisch J.D. and Prentice R.L. (1980). The Statistical Analysis of Failure Time Data. John Wiley & Sons, New York.

Modèle à risques proportionnels de Cox

Utilisez le modèle à risques proportionnels de Cox (ou modèle de Cox) pour modéliser un temps de survie en fonction de variables explicatives quantitatives ou qualitatives. Ce modèle s'intègre dans le cadre des méthodes pour données de survie.

Dans cette section :

Description

Le modèle de Cox est la méthode la plus utilisée dans le cadre de l'analyse des données de survie. Celui-ci permet de modéliser des temps de survie avec des données censurées. Elle est très utilisée dans le domaine médical (temps de survie ou de guérison d'un patient).

Le principe du modèle de Cox est de relier la date d'arrivée d'un évènement à des variables explicatives. Par exemple, dans le domaine médical, on cherche à évaluer l'impact d'un prétraitement sur le temps de guérison d'un patient.

Modèles

Le modèle de Cox se rapproche des modèles de régression classique dans le sens où l'on tente de relier un événement (modélisé par une date) à un certain nombre de variables explicatives.

Le modèle de Cox est considéré comme un modèle semi-paramétrique, il est basé sur l'hypothèse des risques proportionnels.

Le modèle de Cox s'applique à toute situation où l'on étudie le délai de survenue d'un événement. Cet événement peut être la récidive d'une maladie, la réponse à un traitement, le décès, etc. Pour chaque sujet, on connaît la date des dernières nouvelles et l'état par rapport à l'événement étudié.

Les sujets pour lesquels on ne connaît pas l'état à la date de fin de l'étude constituent des données censurées. Les valeurs des variables explicatives X_j sont notées pour chaque sujet à la date de son entrée dans l'étude.

La variable considérée T est le temps écoulé jusqu'à la survenue de l'événement étudié. Le modèle de Cox permet d'exprimer le risque instantané de survenue de l'événement en fonction

de l'instant t et des variables explicatives X_j. Ces variables peuvent représenter des facteurs de risque, des facteurs pronostiques, des traitements, des caractéristiques intrinsèques au sujet,...

Le risque instantané de survenue de l'événement $\lambda(t, X_1, X_2, ..., X_p)$ représente la probabilité d'apparition de l'événement dans un intervalle de temps $[t, t + Dt]$ sachant que l'événement ne s'est pas réalisé avant l'instant t. Le modèle de Cox exprime $\lambda(t, X_1, X_2, ..., X_p)$ sous la forme :

$$\lambda(t, X) = \lambda_0(t) \exp(\beta X)$$

Cette formule appelle quelques commentaires. Le risque instantané se décompose en 2 termes dont l'un dépend du temps t et l'autre des variables X_j. Si, par exemple, les variables X_j représentent des facteurs de risque et si elles sont toutes égales à 0, $\lambda_0(t)$ est le risque instantané de sujets ne présentant aucun facteur de risque. La forme de $\lambda_0(t)$ n'étant pas précisée, c'est plutôt l'association entre les variables X_j et la survenue de l'événement considéré qui est l'intérêt central du modèle. Cela revient à déterminer les coefficients β_j.

Le rapport des risques instantanés de 2 individus dont les caractéristiques respectives sont $(X_1, X_2, ..., X_p)$ et $(X'_1, X'_2, ..., X'_p)$ ne dépend pas du temps. De tels modèles sont dits à risques proportionnels. C'est une hypothèse importante du modèle de Cox.

Si β_j est positif et si 2 sujets ne diffèrent que par la j-ième caractéristique, des valeurs élevées de la j-ième caractéristique sont associées à un risque instantané plus élevé. Inversement si β_j est négatif, des valeurs élevées de la j-ième caractéristique sont associées à un risque instantané plus faible.

Le modèle est estimé en utilisant le principe du maximum de vraisemblance avec quelques modifications, la fonction utilisée est appelée la vraisemblance partielle et a été introduite par Cox (1972). Comme le terme $\lambda_0(t)$ ne nous intéresse pas, il ne sera pas estimé, on minimisera donc une log-vraisemblance partielle :

$$\log\left(L(\beta)\right) = \sum_{i=1}^{n} \beta X_i - \log\left(\sum_{j:t_{(j)} \geq t_{(i)}} \exp\left(\beta X_j\right)\right)$$

Contrairement à la régression linéaire, une solution analytique exacte n'existe pas. Il est donc nécessaire d'utiliser un algorithme itératif. XLSTAT utilise un algorithme de Newton-Raphson. L'utilisateur peut modifier s'il le souhaite le nombre maximum d'itérations et le seuil de convergence.

Les strates

Lorsque l'hypothèse de risques proportionnels n'est pas tenable, il arrive souvent que l'on stratifie le modèle. Si l'hypothèse est tenable sur des sous- échantillons, alors on estime la vraisemblance partielle sur chaque sous- échantillon et on prend la somme des vraisemblances partielles. Dans XLSTAT, les strates doivent être définies par une variable qualitative.

Contraintes pour les variables qualitatives

Le traitement des variables qualitatives se fait en utilisant un tableau disjonctif complet. Néanmoins l'une des modalités de chaque variable doit être supprimée lors de l'estimation pour éviter la dépendance des variables. Dans XLSTAT, on peut choisir de supprimer la première ou la dernière modalité de chaque variable qualitative, ainsi l'effet de la première ou de la dernière modalité correspond à un standard. L'impact des autres modalités se fait relativement à cette modalité omise.

Prise en compte des égalités

Le modèle de Cox a été conçu pour traiter des données de date continues. Néanmoins dans la pratique, il arrive souvent que plusieurs observations se produisent à la même date. Dans ce cas des adaptations de la vraisemblance partielle existent. XLSTAT en propose deux :

La méthode de Breslow (1974) (méthode par défaut) : La vraisemblance a alors la forme suivante :

$$\log\left(L(\beta)\right) = \sum_{i=1}^{T} \beta \sum_{l=1}^{d_i} X_l - d_i \log\left(\sum_{j:t_{(j)} \geq t_{(i)}} \exp\left(\beta X_j\right)\right)$$

où T représente le nombre de dates différentes et d_i est le nombre d'observations au temps $t_{(i)}$.

La méthode d'Efron (1977) : La vraisemblance partielle a alors la forme suivante :

$$\log\left(L(\beta)\right) = \sum_{i=1}^{T} \beta \sum_{l=1}^{d_i} X_l - \sum_{r=0}^{d_i-1} \log\left(\sum_{j:t_{(j)} \geq t_{(i)}} \exp\left(\beta X_j\right) - \frac{r}{d_i} \sum_{j=1}^{d_i} \exp\left(\beta X_j\right)\right)$$

où T représente le nombre de dates différentes et d_i est le nombre d'observations au temps $t_{(i)}$.

Lorsqu'il n'y a pas d'égalité, ces vraisemblances partielles reviennent à la vraisemblance partielle de Cox.

Résidus

Les procédures de diagnostic pour la vérification du modèle sont une partie importante dans un processus de modélisation, et beaucoup de ces procédures sont basées sur les résidus. En analyse de survie, et en particulier lorsque l'on construit un modèle à risque proportionnels de Cox, plusieurs types de résidus sont utilisés à des fins différentes.

Les résidus de Martingale sont utilisés pour examiner la qualité globale de l'ajustement d'un modèle de Cox. Ils sont définis comme suit :

$$M_i = d_i - \Lambda_0(t_i)\exp(x_i\beta)$$

Où :

$$\Lambda_0 = \sum_{t_i<t} \frac{d_i}{\sum_{j\in R_{t(i)}} \exp(x_j\beta)}$$

est la fonction de risque cumulé.

D'après Therneau et al., un problème de ces résidus est, en particulier dans le cas d'un évènement unique dans le modèle de Cox, qu'ils sont biaisés. De plus, ils sont compris entre $-\infty$ et 1.

Pour régler ces problèmes, Therneau et al. ont proposés les résidus de déviance. Ces résidus sont utilisés pour la détection d'observations mal prédites. Ils sont plus symétriques autour de 0 que les résidus de Martingale et sont définis comme suit :

$$D_i = \operatorname{sign}(M_i)\sqrt{-2(M_i + d_i \log(d_i - M_i))}$$

Où M_i est le résidu de Martingale, et sign représente la fonction signe.

Les observations qui possèdent un important résidu de déviance sont celles qui ne sont pas bien adaptées au modèle, et que l'on peut considérer comme des observations aberrantes.

Les résidus de Schoenfeld ont été proposés par Schoenfeld comme étant un résidu partiel essentiel pour vérifier l'hypothèse des risques proportionnels. Schoenfeld a défini ces résidus comme étant la différence entre la valeur observée x_{ik} et son espérance conditionnelle sachant le nombre d'individus R_i encore à risque au temps t_i. Ils s'écrivent :

$$s_{ik} = d_i(x_{ik} - \bar{x_k})$$

Où :

$$\bar{x_k} = \frac{\sum_{j\in R(t_i)} x_{jk}\exp(x\beta)}{\sum_{j\in R(t_i)} \exp(x\beta)}$$

Un graphique de s_{ik} en fonction de t sera centré sur 0 si l'hypothèse de proportionnalité des risques est vérifiée $E(s_{ik}) = 0$.

Les résidus de score sont utilisés pour la détermination d'observations influentes. Ils sont définis comme suit :

$$score_{ik} = d_i(x_{ik} - \bar{x_k}(t_i)) - \sum_{t_n<t_i} (x_{ik} - \bar{x_k}(t_n))\exp(x_i\beta)(\Lambda_0(t_{n-1}))$$

Où

$$\bar{x_k} = \frac{\sum_{j\in R(t_i)} x_{jk}\exp(x\beta)}{\sum_{j\in R(t_i)} \exp(x\beta)}$$

Test de proportionnalité

Une hypothèse fondamentale du modèle à risque proportionnel de Cox est que le rapport de risque soit constant au cours du temps. La violation de cette hypothèse peut entrainer un biais dans l'estimation des coefficients de la régression. Il existe différentes méthodes pour vérifier cette hypothèse.

Dans XLSTAT, il est possible de vérifier cette hypothèse en utilisant les résidus de Schoenfeld. Grambsch et Therneau ont montrés qu'une version normalisée de ces résidus se rapproche de la variation du coefficient de régression au temps k : $E(s_{ik}^*) + \hat{\beta}_k \approx \beta_k(t_i)$.

Où s_{ik}^* est le résidu de Schoenfeld normalisé de la covariable k au temps i : $s_{ik}^* = V^{-1}(\beta, t)S_{ik}$

Et $V(\beta, t)$ est la variance du vecteur des estimations β au temps t.

Ainsi, on peut tester la proportionnalité des prédicteurs en créant une interaction avec le temps ou une transformation du temps (dans XLSTAT, l'estimateur de Kaplan-Meier est utilisé comme transformation du temps). Pour chaque variable, on va tester la nullité de l'espérance des résidus de Schoenfeld : $H_0 : \beta_j(t) = \beta_j$ contre $H_1 : \beta_j(t) \neq \beta_j$

Une p-valeur < α indique une violation de l'hypothèse de proportionnalité.

Le test global permet de vérifier cette hypothèse pour l'ensemble du modèle : $H_0 : \beta(t) = \beta$ contre $H_1 : \beta(t) \neq \beta$

Indices de validation du modèle

XLSTAT-Life vous permet d'afficher des indices de validation du modèle. Ceux- ci sont obtenus en utilisant la méthode du bootstrap.

Ainsi XLSTAT-Life propose pour chaque indice, la valeur moyenne, l'écart-type ainsi qu'un intervalle de confiance associé à l'indice.

Ces indices sont :

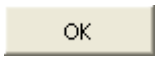
- R^2(Cox et Snell) : coefficient compris comme le R^2 entre 0 et 1 qui mesure le bon ajustement du modèle. Ce coefficient est égal à 1 moins le rapport de la vraisemblance du modèle ajusté sur la vraisemblance du modèle indépendant ;
- R^2(Nagelkerke) : coefficient compris comme le R^2 entre 0 et 1 qui mesure le bon ajustement du modèle. Ce coefficient est égal au rapport du R^2 de Cox et Snell, divisé par 1 moins la vraisemblance du modèle indépendant
- L'indice de Shrinkage : Cet indice permet de quantifier l'overfitting du modèle. Lorsque celui-ci est plus petit que 0,85, on dira qu'il y a overfitting dans le modèle et qu'il faut réduire le nombre de paramètres du modèle.

- L'indice c : L'indice de concordance ou indice général de discrimination permet d'évaluer la qualité prédictive du modèle. Lorsqu'il est proche de 1, cette qualité est bonne lorsqu'il est proche de 0, elle est mauvaise.
- Le D de Sommer : cet indice est directement lié au précédent, on a $D = 2 * (c - 0,5)$. Il se comporte comme une corrélation et varie de -1 à 1.

Ils permettent à l'utilisateur de valider le modèle de Cox obtenu. Pour une description détaillée des processus de bootstrap et de validation du modèle de Cox, on peut voir Harrell et al. (1996).

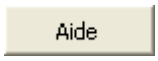
Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.



: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas (mode colonnes), XLSTAT considère que les sites sont en lignes et les objets/variables en colonnes. Si la flèche est vers la droite (mode lignes), XLSTAT considère que les objets/variables sont en lignes et les sites en colonnes.

Onglet **Général** :

Données de dates : sélectionnez les données correspondant aux dates auxquelles se produisent les événements ou les censures. Si un en-tête a été sélectionné sur la première ligne, veillez à ce que l'option « libellés des colonnes » soit activée.

Indicateur d'état : sélectionnez ici les données correspondant à une « donnée événement » ou à une « donnée censurée ». Si un en-tête a été sélectionné sur la première ligne, veillez à ce que l'option « libellés des colonnes » soit activée

Code événement : entrez le code utilisé pour identifier une « donnée événement ». La valeur par défaut est 1.

Code censuré : entrez ici le code utilisé pour identifier une « donnée censurée ». La valeur par défaut est 0.

Variables explicatives :

Quantitatives : activez cette option si vous voulez inclure une ou plusieurs variables explicatives quantitatives dans le modèle. Sélectionnez alors la ou les variables correspondantes sur la feuille Excel. Les données sélectionnées doivent être de type numérique. Si le libellé des variables a été sélectionné, veuillez vérifier que l'option « Libellés des colonnes » est activée.

Qualitatives : activez cette option si vous voulez inclure une ou plusieurs variables explicatives qualitatives dans le modèle. Sélectionnez alors la ou les variables correspondantes sur la feuille Excel. Les données sélectionnées peuvent être de tout type, mais les données numériques sont automatiquement considérées comme nominales. Si le libellé des variables a été sélectionné, veuillez vérifier que l'option « Libellés des colonnes » est activée (voir *description*).

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des colonnes : activez cette option si la première ligne des données sélectionnées (variables de temps, de censure et explicatives) contient un libellé.

Strates : activez cette option puis sélectionnez ici les données d'appartenance à des strates (voir *description*).

Onglet **Options** :

Niveau de signification (%) : entrez la valeur du niveau de signification à utiliser pour les tests (valeur par défaut : 5%). Cette valeur est aussi utilisée pour déterminer les intervalles de confiance pour les statistiques calculées.

Prise en charge des égalités : sélectionnez la méthode permettant d'estimer le modèle lorsque plusieurs observations se produisent à la même date (voir la section description). Méthode par défaut : méthode de Breslow.

Interactions / Niveau : activez cette option pour inclure des interactions dans le modèle puis entrez le niveau maximum d'interaction (valeur comprise entre 1 et 4).

Conditions d'arrêt :

- **Itérations** : entrez le nombre maximal d'itérations pour l'algorithme de Newton-Raphson. Les calculs sont interrompus dès que le nombre maximal d'itérations est dépassé. Valeur par défaut : 100.
- **Convergence** : entrez la valeur seuil d'évolution maximale de log vraisemblance d'une itération à l'autre, qui une fois atteinte permet de considérer que l'algorithme a convergé. Valeur par défaut : 0,000001.

Sélection du modèle : activez cette option si vous souhaitez utiliser l'une des deux méthodes de sélection proposées :

- **Ascendante** : le processus de sélection commence par l'ajout de la variable ayant la plus forte contribution au modèle. Si une seconde variable est telle que sa probabilité d'entrée est supérieure à la **valeur seuil pour entrer**, alors elle est ajoutée au modèle.
- **Descendante** : cette méthode est similaire à la précédente, mais part d'un modèle complet.

Onglet **Données manquantes** :

Ne pas accepter les valeurs manquantes : activez cette option pour que XLSTAT empêche la poursuite des calculs si des valeurs manquantes sont détectées.

Supprimer les observations : activez cette option pour supprimer les observations comportant des données manquantes.

Estimer les données manquantes : activez cette option pour estimer les données manquantes avant le début des calculs.

- **Moyenne ou mode** : activez cette option pour estimer les données manquantes en utilisant la moyenne (variables quantitatives) ou le mode (variables qualitatives) pour les variables correspondantes.
- **Plus proche voisin** : activez cette option pour estimer les données manquantes d'une observation en recherchant le plus proche voisin de l'observation.

Onglet **Sorties**:

Statistiques descriptives : activez cette option pour afficher les statistiques descriptives pour les variables sélectionnées.

Coefficients d'ajustement : activez cette option pour afficher le tableau des statistiques d'ajustement du modèle.

Test de l'hypothèse nulle $H_0 : \beta = 0$: activez cette option pour afficher le tableau des statistiques afin de tester l'hypothèse nulle H_0 (rapport de vraisemblance, critère de Wald et critère du score).

Coefficients du modèle : activez cette option pour afficher le tableau des coefficients du modèle. Les trois dernières colonnes donnent les rapports de risque et leurs intervalles de confiance (le rapport de risque est l'exponentielle du coefficient).

Test de proportionnalité : activez cette option pour afficher les résultats du test de proportionnalité.

Prédictions : activez cette option pour afficher le vecteur des prédictions. On calcule ces prédicteurs linéaires comme suit : $(x_i - mean(x_i))\beta_i$.

Résidus : activez cette option pour afficher les différents types de résidus pour l'ensemble des observations (résidus de déviance, résidus de martingale, résidus de Schoenfeld et résidus du Score) (Voir partie description de ce chapitre).

Statistiques rééchantillonnées : activez cette option afin d'afficher les indices de validation obtenus en utilisant la méthode du bootstrap (voir partie description de ce chapitre).

- **Rééchantillonnages** : si l'option précédente est activée, entrez le nombre d'échantillons à générer lorsque la méthode bootstrap est utilisée.

Onglet **Graphiques** :

Fonction de survie cumulée : activez cette option pour afficher les graphiques relatifs à la fonction de survie cumulée.

-Log(FSC) : activez cette option si vous souhaitez que XLSTAT affiche le –Log() de la fonction de survie (FSC).

Log(-Log(FSC)) : activez cette option si vous souhaitez que XLSTAT affiche le Log(–Log()) de la fonction de survie (FSC).

Fonction de risque : activez cette option si vous souhaitez que XLSTAT affiche la fonction de risque à la moyenne des variables explicatives.

Résidus : activez cette option si vous souhaitez que XLSTAT affiche le graphique des résidus en fonction du temps.

Résultats

XLSTAT propose un nombre important de tableaux et de graphiques afin de faciliter l'analyse et l'interprétation des résultats.

Statistiques descriptives : le tableau de statistiques descriptives présente pour toutes les variables sélectionnées des statistiques simples. Pour les variables quantitatives, sont affichés le nombre d'observations, le nombre de données manquantes, le nombre de données non manquantes, la moyenne, et l'écart-type (non biaisé). Pour les variables qualitatives sont affichées les modalités leurs effectifs et pourcentage respectifs.

Synthèse de la sélection des variables : dans le cas où une méthode de sélection a été choisie, XLSTAT affiche la synthèse de la sélection. Dans le cas d'une sélection pas à pas, les statistiques correspondant aux différentes étapes sont affichées.

Coefficients d'ajustement : dans ce tableau est affichée une série de statistiques pour le modèle indépendant (correspondant au cas où il n'y aurait aucune variables dans le modèle) et pour le modèle ajusté.

- **Observations** : le nombre total d'observations prises en compte ;
- **DDL** : degrés de liberté ;
- **-2 Log(Vrais.)** : le logarithme de la fonction de vraisemblance associée au modèle ;
- **AIC** : le critère d'information d'Akaike (Akaike's Information Criterion) ;
- **SBC** : le critère bayésien de Schwarz (Schwarz's Bayesian Criterion) ;
- **Itérations** : nombre d'itérations nécessaires à la convergence de l'algorithme.

Test de l'hypothèse nulle H_0 : $\beta = 0$: l'hypothèse H_0 correspond au modèle indépendant (sans variables explicatives) ; on cherche à vérifier si le modèle ajusté est significativement plus performant que ce modèle. Trois tests sont proposés : le test du rapport des vraisemblance (-2 Log(Vrais.)), le test du Score, et le test de Wald. Les trois statistiques suivent une loi du χ^2 dont les degrés de liberté sont indiqués.

Paramètres du modèle : pour chaque variable sont affichés l'estimation du paramètre, l'écart-type correspondant, le χ^2 de Wald, la p-value correspondante. Par ailleurs, le rapport de risque (exponentielle du coefficient) est donné ainsi qu'un intervalle de confiance associé.

Test de proportionnalité : Pour chaque variable sont affichés la corrélation entre les résidus de Schoenfeld et le vecteur de temps (1 – Kaplan Meier), la statistique de test et sa p-valeur.

Les **prédictions** sont données pour chaque observation.

Les **résidus** sont donnés pour chaque observation.

Exemple

Un exemple de modèle de Cox est disponible sur le Centre d'aide XLSTAT :

http://www.xlstat.com/demo-coxf.htm

Bibliographie

Cox D. R. (1972). Regression Models and Life Tables (with Discussion). *Journal of the Royal Statistical Society, Series B* 34:187-220.

Breslow N. E. (1974). Covariance analysis of censored survival data. *Biometrics*, 30:89-99.

Efron B. (1977). Efficiency of Cox's likelihood function for censored data. *Journal of the American Statistical Association*, 72:557-565

Schoenfeld D. (1982). Partial residuals for the proportional hazards regression model, Biometrika, 69, 239 - 241.

Cox D. R. and Oakes D. (1984). Analysis of Survival Data. Chapman and Hall, London.

Therneau T. M., Grambsch P. M., and Fleming T. R. (1990). Martingale- based residuals for survival models, Biometrika. 77, 147 - 160.

Grambsch P. M. and Therneau T. M. (1994). Proportional hazards tests and diagnostics based on weighted residuals, Biometrika, 81, 515 – 526.

Collett D. (1994). Modeling Survival Data In Medical Research. Chapman and Hall, London.

Harrell F.E. Jr., Lee K.L. and Mark D.B. (1996). Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy and measuring and reducing errors. *Statistics in Medicine*, **15**, 361-387

Hill C., Com-Nougué C., Kramar A., Moreau T., O'Quigley J. Senoussi R. and Chastang C. (1996). Analyse statistique des données de survie. 2nd Edition, INSERM, Médecine-Sciences, Flammarion.

Kalbfleisch J. D. and Prentice R. L. (2002). The Statistical Analysis of Failure Time Data. 2nd edition, John Wiley & Sons, New York.

Modèle à risques proportionnels avec données censurées par intervalle

Utilisez le modèle à risques proportionnels avec données censurées par intervalle pour modéliser un temps de survie en fonction de variables explicatives quantitatives ou qualitatives. Ce modèle s'intègre dans le cadre des méthodes pour données de survie.

Dans cette section :

Description

Le modèle à risques proportionnels est un des modèles de régression les plus populaires pour modéliser les temps de survie avec des données censurées.

Le principe de ce modèle est de relier la date d'arrivée d'un évènement à des variables explicatives. Par exemple, dans le domaine médical, on cherche à évaluer l'impact d'un prétraitement sur le temps de guérison d'un patient.

Le 1er modèle à risques proportionnels, introduit par Cox en 1972, fonctionne avec des données non censurées et des données censurées à droite. Le but de ce modèle à risques proportionnels avec données censurées par intervalle est donc le même que pour le modèle de Cox, mais il va aussi être possible de modéliser des temps de survies pour des données censurées par intervalle, des données non censurées, censurées à gauche ou censurées à droite.

Dans le cas où les données ne contiennent que des observations non censurées et censurées à droite, il est possible, avec cette fonction de reproduire les résultats d'un modèle de Cox. Cependant, il est recommandé d'utiliser le modèle à risques proportionnels de Cox, car il fournit une méthode plus adaptée dans ce genre de cas.

Modèles

Vraisemblance

Dans un modèle à risques proportionnels, le temps de survie de chaque individu d'une population est supposé suivre sa propre fonction de risque $\lambda_i(t)$, définie par :

$$\lambda_i(t) = \lambda_0(t)\exp(X_i'\beta)$$

Où $\lambda_0(t)$ est la fonction de risque de base, X_i est le vecteur des variables explicatives pour le i-ème individu, et β est le vecteur des coefficients de régression.

Dans ce modèle, on va supposer que les observations à analyser sont censurées par intervalle :

$$\{(L_i, R_i], X_i\}_{i=1}^n$$

Où L_i et $R_i$représentent les limites gauche et droite de l'intervalle de censure de l'individu i avec $L_i \leq R_i$. On peut noter que si $L_i = R_i$ le temps de l'individu i est non censuré, si $L_i = 0$ le temps de l'individu i est censuré à gauche, et si $R_i = \infty$ le temps de l'individu i est censuré à droite.

On note ensuite $F(t|X_i)$ la fonction de répartition de l'individu i au temps t. Sous le modèle à risques proportionnels, elle est donnée par :

$$F(t|X_i) = 1 - \exp(-\Lambda_0(t)\exp(X_i'\beta))$$

Où $\Lambda_0(t)$ est la fonction cumulée du risque de base et β les coefficients de régression.

On suppose que le temps entre le début de l'étude et l'évènement est indépendant du processus d'observation. Sous cette hypothèse, la fonction de vraisemblance est donnée par :

$$L = \prod_{i=1}^n \{F(R_i|X_i) - F(L_i|X_i)\}$$

Si on distingue maintenant les trois types de censures, on obtient :

$$L = \prod_{i=1}^n \{F(R_i|X_i)^{\delta_1}\{F(R_i|X_i) - F(L_i|X_i)\}^{\delta_2}\{1 - F(L_i|X_i)\}^{\delta_3}$$

Où δ_1(resp δ_2, δ_3) = 1 si il y a censure à gauche (resp par intervalle, à droite) et 0 sinon.

Spline cubique

Dans la fonction de vraisemblance donnée précédemment, les paramètres inconnus sont les coefficients de régression β ainsi que la fonction cumulée du risque de base $\Lambda_0(t)$. On va donc chercher à estimer $\Lambda_0(t)$ à l'aide de I-splines (Ramsay, 1988). Cette approche nous donne la représentation suivante :

$$\Lambda_0(t) = \sum_{l=1}^k \gamma_l b_l(t)$$

Où $b_l(t)$ sont les fonctions de base des I-splines, et γ_l sont des coefficients non nuls qui garantissent que $\Lambda_0(t)$ est non décroissant.

Il existe différents paramètres à spécifier pour construire ces fonctions de base des I-splines : le degré des splines ainsi que le nombre et le placement des nœuds auxquels on estime ces fonctions. Dans le cadre de cette fonction, l'utilisation de splines cubique (ordre = 3) a été privilégié. En ce qui concerne le nombre de nœuds, il est fixé à 3, espacés régulièrement. Il est aussi possible de faire le choix d'optimiser ce nombre de nœud. Dans ce cas, on va construire un modèle pour différents nombres de nœuds et en calculer l'AIC, le modèle final sera celui dont l'AIC est le plus faible. Où l'AIC est le critère d'information d'Akaike (Akaike's Information Criterion).

Algorithme EM

L'algorithme EM (Expectation-Maximisation) est un algorithme itératif. C'est une méthode d'estimation paramétrique s'inscrivant dans le cadre du maximum de vraisemblance.

Il est composé de deux étapes. La première consiste à calculer l'espérance du logarithme de la vraisemblance du modèle en fonction des données observées et de l'estimation des paramètres $\theta^{(d)} = (\beta^{(d)}, \gamma^{(d)})$. Ce qui donne :

$$Q(\theta, \theta^{(d)}) = E[\log(L(\theta))]$$

La seconde étape consiste à optimiser cette fonction. Après calculs, on peut voir que $\theta^{(d+1)}$ est une solution au système d'équations donné par :

$$\frac{\partial Q(\theta, \theta^{(d)})}{\partial \beta} = 0$$

Et

$$\frac{\partial Q(\theta, \theta^{(d)})}{\partial \gamma_l} = 0$$

Où $l = 1, \ldots, k$.

En résolvant la 2ème équation, on obtient directement une expression pour $\gamma^{(d+1)}$ en termes de $\beta^{(d+1)}$ et des données observées pour chaque l. Ainsi, en remplaçant γ_l dans $\frac{\partial Q(\theta, \theta^{(d)})}{\partial \beta} = 0$ par l'expression de $\gamma^{(d+1)}$, on peut obtenir $\beta^{(d+1)}$, ce qui permet ensuite le calcul direct de $\gamma_l^{(d+1)} = \gamma_l^{(d)}(\beta^{(d+1)})$.

La résolution du système d'équation $\frac{\partial Q(\theta, \theta^{(d)})}{\partial \beta} = 0$ est, dans cette fonction, réalisée à l'aide de l'algorithme de Nelder-Mead, qui est un algorithme d'optimisation.

Estimation de la variance

Pour calculer la matrice d'information de Fisher $\big(I(\theta)\big)$, la méthode de Louis (Louis, 1982) est utilisée. La matrice de variance-covariance de $\hat{\theta}$ sera alors donnée en inversant cette matrice d'informations de Fisher. La matrice d'information de Fisher est donnée par :

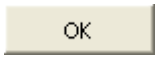
$$I(\theta) = -\frac{\partial^2 Q(\theta, \hat{\theta})}{\partial\theta\partial\hat{\theta}} - var(\frac{\partial \log(L(\theta))}{\partial\theta})$$

Contraintes pour les variables qualitatives

Le traitement des variables qualitatives se fait en utilisant un tableau disjonctif complet. Néanmoins l'une des modalités de chaque variable doit être supprimée lors de l'estimation pour éviter la dépendance des variables. Dans XLSTAT, on peut choisir de supprimer la première ou la dernière modalité de chaque variable qualitative, ainsi l'effet de la première ou de la dernière modalité correspond à un standard. L'impact des autres modalités se fait relativement à cette modalité omise.

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas (mode colonnes), XLSTAT considère que les sites sont en lignes et les objets/variables en colonnes. Si la flèche est vers la droite (mode lignes), XLSTAT considère que les objets/variables sont en lignes et les sites en colonnes.

Onglet **Général** :

Borne gauche : sélectionnez ici les données correspondant à la borne gauche de l'intervalle. Si la donnée est censurée à gauche, mettez une valeur de 0. Si la donnée est censurée par intervalle, cette valeur doit être inférieure à celle de la borne droite.

Borne droite : sélectionnez ici les données correspondant à la borne droite de l'intervalle. Si la donnée est censurée à droite, mettez une valeur de 0. Si la donnée est censurée par intervalle, cette valeur doit être supérieure à celle de la borne gauche.

Variables explicatives :

Quantitatives : activez cette option si vous voulez inclure une ou plusieurs variables explicatives quantitatives dans le modèle. Sélectionnez alors la ou les variables correspondantes sur la feuille Excel. Les données sélectionnées doivent être de type numérique. Si le libellé des variables a été sélectionné, veuillez vérifier que l'option « Libellés des colonnes » est activée.

Qualitatives : activez cette option si vous voulez inclure une ou plusieurs variables explicatives qualitatives dans le modèle. Sélectionnez alors la ou les variables correspondantes sur la feuille Excel. Les données sélectionnées peuvent être de tout type, mais les données numériques sont automatiquement considérées comme nominales. Si le libellé des variables a été sélectionné, veuillez vérifier que l'option « Libellés des colonnes » est activée (voir description).

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des colonnes : activez cette option si la première ligne des données sélectionnées (variables de temps, de censure et explicatives) contient un libellé.

Code de censure : Sélectionnez ici le vecteur des codes de censure correspondants aux valeurs des codes rentrés ci-après.

Code non censure : entrez le code utilisé pour identifier une donnée non censurée. La valeur par défaut est 0.

Code censure gauche : entrez le code utilisé pour identifier une donnée censurée à gauche. La valeur par défaut est 1.

Code censure intervalle : entrez le code utilisé pour identifier une donnée censurée par intervalle. La valeur par défaut est 2.

Code censure droite : entrez le code utilisé pour identifier une donnée censurée à droite. La valeur par défaut est 3.

Onglet **Options** :

Niveau de signification (%) : entrez la valeur du niveau de signification à utiliser pour les tests (valeur par défaut : 5%). Cette valeur est aussi utilisée pour déterminer les intervalles de confiance pour les statistiques calculées.

Optimiser le nombre de nœuds : activez cette option pour optimiser le nombre de nœuds utilisé pour le calcul des splines. Le « meilleur » nombre de nœuds sera alors celui qui optimise l'AIC du modèle. Dans le cas où cette option n'est pas activée, le nombre de nœuds sera alors de 3.

Interactions / Niveau : activez cette option pour inclure des interactions dans le modèle puis entrez le niveau maximum d'interaction (valeur comprise entre 1 et 4).

Conditions d'arrêt :

- **Itérations** : entrez le nombre maximal d'itérations pour l'algorithme de Newton-Raphson. Les calculs sont interrompus dès que le nombre maximal d'itérations est dépassé. Valeur par défaut : 100.
- **Convergence** : entrez la valeur seuil d'évolution maximale de log vraisemblance d'une itération à l'autre, qui une fois atteinte permet de considérer que l'algorithme a convergé. Valeur par défaut : 0,000001.

Onglet **Données manquantes** :

Ne pas accepter les valeurs manquantes : activez cette option pour que XLSTAT empêche la poursuite des calculs si des valeurs manquantes sont détectées.

Supprimer les observations : activez cette option pour supprimer les observations comportant des données manquantes.

Estimer les données manquantes : activez cette option pour estimer les données manquantes avant le début des calculs.

- **Moyenne ou mode** : activez cette option pour estimer les données manquantes en utilisant la moyenne (variables quantitatives) ou le mode (variables qualitatives) pour les variables correspondantes.
- **Plus proche voisin** : activez cette option pour estimer les données manquantes d'une observation en recherchant le plus proche voisin de l'observation.

Onglet **Sorties** :

Statistiques descriptives : activez cette option pour afficher les statistiques descriptives pour les variables sélectionnées.

Coefficients d'ajustement : activez cette option pour afficher le tableau des statistiques d'ajustement du modèle.

Coefficients du modèle : activez cette option pour afficher le tableau des coefficients du modèle. Les trois dernières colonnes donnent les rapports de risque et leurs intervalles de confiance (le rapport de risque est l'exponentielle du coefficient).

Prédictions : activez cette option pour afficher le vecteur des prédictions. On calcule ces prédicteurs linéaires comme suit : $(x_i - mean(x_i))\beta_i$.

Onglet **Graphiques** :

Fonction de survie cumulée : activez cette option pour afficher les graphiques relatifs à la fonction de survie cumulée.

-Log(FSC) : activez cette option si vous souhaitez que XLSTAT affiche le –Log() de la fonction de survie (FSC).

Log(-Log(FSC)) : activez cette option si vous souhaitez que XLSTAT affiche le Log(–Log()) de la fonction de survie (FSC).

Résultats

XLSTAT propose un nombre important de tableaux et de graphiques afin de faciliter l'analyse et l'interprétation des résultats.

Statistiques descriptives : le tableau de statistiques descriptives présente pour toutes les variables sélectionnées des statistiques simples. Pour les variables quantitatives, sont affichés le nombre d'observations, le nombre de données manquantes, le nombre de données non manquantes, la moyenne, et l'écart-type (non biaisé). Pour les variables qualitatives sont affichées les modalités leurs effectifs et pourcentage respectifs.

Coefficients d'ajustement : dans ce tableau est affichée une série de statistiques pour le modèle indépendant (correspondant au cas où il n'y aurait aucune variables dans le modèle) et pour le modèle ajusté.

- **Observations** : le nombre total d'observations prises en compte ;
- **DDL** : degrés de liberté ;
- **-2 Log(Vrais.)** : le logarithme de la fonction de vraisemblance associée au modèle ;
- **AIC** : le critère d'information d'Akaike (Akaike's Information Criterion) ;
- **SBC** : le critère bayésien de Schwarz (Schwarz's Bayesian Criterion) ;
- **Itérations** : nombre d'itérations nécessaires à la convergence de l'algorithme.

Paramètres du modèle : pour chaque variable sont affichés l'estimation du paramètre, l'écart-type correspondant, le χ^2 de Wald, la p-value correspondante. Par ailleurs, le rapport de risque (exponentielle du coefficient) est donné ainsi qu'un intervalle de confiance associé.

Les **prédictions** sont données pour chaque observation.

Exemple

Un exemple de modèle de modèle à risques proportionnels avec données censurées par intervalle est disponible sur le Centre d'aide XLSTAT à l'adresse

http://www.xlstat.com/demo-coif.htm

Bibliographie

Cox D. R. (1972). Regression Models and Life Tables (with Discussion). Journal of the Royal Statistical Society, Series B 34:187-220.

Wang L., McMahan C. S., Hudgens M. G., Qureshi Z. P. (2016). A Flexible, Computationally Efficient Method for Fitting the Proportional Hazards Model to Interval-Censored Data. Biometrics 72, 222-231.

McMahan C. S., Wang L., Tebbs J. M. (2013). Re gression analysis for current status data using the EM algorithm. Statistics in medicine, Vol. 32(25), 4452-4466.

Rosenberg P. S. (1995). Hazard function estimation using B-Splines. Biometrics 51, 874-887.

Ramsay J. O. (1988). Monotone regression splines in action. Statistical Science, Vol. 3, No. 4, 425-461.

Nash J. C. (1979). Compact numerical methods for computers: linear algebra and function minimisation.

Modèles de survie paramétriques (modèle de Weibull)

Cet outil permet d'appliquer les modèles de survie paramétriques. Il permet aussi bien d'obtenir des courbes de survie paramétriques que des régressions de survie paramétriques. Les modèles de survie paramétriques permettent de modéliser un temps de survie en fonction d'une distribution et, si nécessaire, de variables explicatives quantitatives ou qualitatives. Ces modèles s'intègrent dans le cadre des méthodes d'analyse de données de survie.

Dans cette section :

Description

Le modèle de survie paramétrique est une méthode s'appliquant dans le cadre de l'analyse des données de survie. Celui-ci permet de modéliser des temps de survie avec des données censurées. Il est très utilisé dans le domaine médical (temps de survie ou de guérison d'un patient).

Le principe du modèle de survie paramétrique est de relier la date de survenance d'un évènement à une distribution de probabilité (on utilise souvent la distribution de Weibull) et, lorsque cela est nécessaire, à des variables explicatives. Par exemple, dans le domaine médical, on cherche à évaluer l'impact d'un prétraitement sur le temps de guérison d'un patient, en supposant que ce temps suit une distribution de Weibull.

XLSTAT-Life propose deux outils pour les modèles de survie paramétriques :

- La régression paramétrique sur données de survie qui permet d'appliquer un modèle de régression et donc d'analyser l'impact de variables explicatives sur le temps de survie (en supposant toujours une distribution sous-jacente).
- Les courbes de survie paramétriques qui permettent de modéliser le temps de survie en fonction de la distribution choisie.

Ces deux méthodes sont exactement équivalentes d'un point de vue méthodologique, la différence réside dans le fait que, dans le premier cas, on a des variables explicatives alors que dans le second cas nous n'en diposons pas.

Modèles

Le modèle de survie paramétrique se rapproche des modèles de régression classique dans le sens où l'on tente de relier un événement (modélisé par une date) à un certain nombre de variables explicatives.

Le modèle de survie paramétrique est un modèle paramétrique, il est basé sur l'hypothèse que les temps de survie suivent une distribution fixée a priori. On suppose donc une structure pour la fonction de risque qui est associée à la distribution choisie.

Le modèle de survie paramétrique s'applique à toute situation où l'on étudie le délai de survenance d'un événement. Cet événement peut être la récidive d'une maladie, la réponse à un traitement, le décès, etc. Pour chaque sujet, on connaît la date des dernières nouvelles et l'état par rapport à l'événement étudié.

Les sujets pour lesquels on ne connaît pas l'état à la date de fin de l'étude constituent des données censurées. Les valeurs des variables explicatives X_j sont notées pour chaque sujet à la date de son entrée dans l'étude.

La variable considérée T est le temps écoulé jusqu'à la survenance de l'événement étudié. Le modèle de survie paramétrique permet d'exprimer le risque de survenance de l'événement en fonction de l'instant t et des variables explicatives X_j. Ces variables peuvent représenter des facteurs de risque, des facteurs pronostiques, des traitements, des caractéristiques intrinsèques au sujet, etc.

La fonction de survie, notée $S(t)$, est définie en fonction de la distribution choisie. XLSTAT-Life propose différentes distributions, parmi lesquelles, la distribution exponentielle (son taux de survie est constant, $h(t) = l$), la distribution de Weibull (qui est souvent appelée modèle de Weibull), les distributions des valeurs extrême.

Les modèles exponentiel et de Weibull sont extrêmement intéressant car ce sont en même temps des modèles à risques proportionnels (comme le modèle de Cox) et des modèles de temps de sortie accéléré (pour tout individus i et j avec des temps de survie $S_i()$ et $S_j()$, il existe une constante ϕ tel que $S_i(t) = S_j(\phi * t)$ pour tout t).

L'estimation de ce type de modèles se fait par la méthode du maximum de vraisemblance. On utilise généralement comme variable à expliquer $Y = log(T)$ (c'est le cas des modèles exponentiels et de Weibull).

Contrairement à la régression linéaire, une solution analytique exacte n'existe pas. Il est donc nécessaire d'utiliser un algorithme itératif. XLSTAT utilise un algorithme de Newton-Raphson. L'utilisateur peut modifier s'il le souhaite le nombre maximum d'itérations et le seuil de convergence.

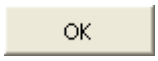
L'interprétation des résultats se fait à la fois en étudiant les graphiques associés aux fonctions de survie cumulée et en étudiant les tableaux des coefficients et des indices de qualité d'ajustement.

Contraintes pour les variables qualitatives

Le traitement des variables qualitatives se fait en utilisant un tableau disjonctif complet. Néanmoins l'une des modalités de chaque variable doit être supprimée lors de l'estimation pour éviter la dépendance des variables. Dans le cadre d'XLSTAT, c'est la première ou la dernière modalité de chaque variable qualitative qui est supprimée, ainsi l'effet de la première modalité correspond à un standard. L'impact des autres modalités se fait relativement à cette première modalité omise.

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas (mode colonnes), XLSTAT considère que les sites sont en lignes et les objets/variables en colonnes. Si la flèche est vers la droite (mode lignes), XLSTAT considère que les objets/variables sont en lignes et les sites en colonnes.

Onglet **Général** :

Données de dates : sélectionnez les données correspondant aux dates auxquelles se produisent les événements ou les censures. Si un en-tête a été sélectionné sur la première ligne, veillez à ce que l'option « libellés des colonnes » soit activée.

Indicateur d'état : sélectionnez ici les données correspondant à une « donnée événement » ou à une « donnée censurée ». Si un en-tête a été sélectionné sur la première ligne, veillez à ce que l'option « libellés des colonnes » soit activée

Code événement : entrez le code utilisé pour identifier une « donnée événement ». La valeur par défaut est 1.

Code censuré : entrez ici le code utilisé pour identifier une « donnée censurée ». La valeur par défaut est 0.

Variables explicatives (dans le cas du modèle de régression paramétrique) :

Quantitatives : activez cette option si vous voulez inclure une ou plusieurs variables explicatives quantitatives dans le modèle. Sélectionnez alors la ou les variables correspondantes sur la feuille Excel. Les données sélectionnées doivent être de type numérique. Si le libellé des variables a été sélectionné, veuillez vérifier que l'option « Libellés des colonnes » est activée.

Qualitatives : activez cette option si vous voulez inclure une ou plusieurs variables explicatives qualitatives dans le modèle. Sélectionnez alors la ou les variables correspondantes sur la feuille Excel. Les données sélectionnées peuvent être de tout type, mais les données numériques sont automatiquement considérées comme nominales. Si le libellé des variables a été sélectionné, veuillez vérifier que l'option « Libellés des colonnes » est activée (voir *description*).

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des colonnes : activez cette option si la première ligne des données sélectionnées (variables de temps, de censure et explicatives) contient un libellé.

Distribution : choisissez la distribution que vous voulez appliquer à votre modèle (voir *description*).

Poids dans la régression : activez cette option si vous voulez pondérer l'influence des observations pour l'ajustement du modèle. Si vous n'activez pas cette option, les poids seront tous considérés comme valant 1. Les poids doivent être impérativement supérieurs ou égaux à 0. Si un en-tête de colonne a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

Onglet **Options** :

Niveau de signification (%) : entrez la valeur du niveau de signification à utiliser pour les tests (valeur par défaut : 5%). Cette valeur est aussi utilisée pour déterminer les intervalles de confiance pour les statistiques calculées.

Constante fixée : activez cette option pour fixer la constante du modèle de régression à une valeur que vous devez ensuite saisir (0 par défaut).

Tolérance : activez cette option pour permettre à l'algorithme de calcul de ne pas prendre en compte les variables qui seraient soit constantes soit trop corrélées avec d'autres variables déjà utilisées dans le modèle (0.0001 par défaut).

Valeurs de départ : activez cette option pour donner un point de départ à XLSTAT. Sélectionnez alors les cellules correspondant aux valeurs initiales des paramètres. Le nombre de lignes sélectionnées doit correspondre au nombre de paramètres.

Contraintes : des détails sur les différentes options sont disponibles dans la section description.

a1 = 0 : choisissez cette option pour que le paramètre de la première modalité de chaque facteur soit fixé à 0.

an = 0 : choisissez cette option pour que le paramètre de la dernière modalité de chaque facteur soit fixé à 0.

Conditions d'arrêt :

- **Itérations** : entrez le nombre maximal d'itérations pour l'algorithme de Newton-Raphson. Les calculs sont interrompus dès que le nombre maximal d'itérations est dépassé. Valeur par défaut : 100.
- **Convergence** : entrez la valeur seuil d'évolution maximale de log vraisemblance d'une itération à l'autre, qui une fois atteinte permet de considérer que l'algorithme a convergé. Valeur par défaut : 0,000001.

Sélection du modèle : activez cette option si vous souhaitez utiliser l'une des deux méthodes de sélection proposées :

- **Ascendante** : le processus de sélection commence par l'ajout de la variable ayant la plus forte contribution au modèle. Si une seconde variable est telle que sa probabilité d'entrée est supérieure à la **valeur seuil pour entrer**, alors elle est ajoutée au modèle.
- **Descendante** : cette méthode est similaire à la précédente, mais part d'un modèle complet.

Onglet **Données manquantes** :

Ne pas accepter les valeurs manquantes : activez cette option pour que XLSTAT empêche la poursuite des calculs si des valeurs manquantes sont détectées.

Supprimer les observations : activez cette option pour supprimer les observations comportant des données manquantes.

Onglet **Sorties** :

Statistiques descriptives : activez cette option pour afficher les statistiques descriptives pour les variables sélectionnées.

Coefficients d'ajustement : activez cette option pour afficher le tableau des statistiques d'ajustement du modèle.

Test de l'hypothèse nulle H0 : beta=0 : activez cette option pour afficher le tableau des statistique afin de tester l'hypothèse nulle H0 (rapport de vraisemblance, critère de Wald et critère du score).

Coefficients du modèle : activez cette option pour afficher le tableau des paramètres du modèle.

Résidus et prédictions : activez cette option pour afficher les différents types de résidus pour l'ensemble des observations (résidus standardisés et résidus de Cox-Snell). Dans ce tableau sont aussi affichées les valeurs prédites pour la fonction de répartition cumulée.

Quantiles : activez cette option pour afficher les quantiles prédits pour chaque observation (dans le cas de la régression) et pour l'ensemble de la courbe de survie (dans le cas des courbes paramétriques). XLSTAT-Life donne les quantiles à 1, 5, 10, 25, 50, 75, 90, 95, et 99 %.

Onglet **Graphiques** :

Fonction de survie cumulée : activez cette option pour afficher les graphiques relatifs à la fonction de survie cumulée.

-Log(FSC) : activez cette option si vous souhaitez que XLSTAT affiche le –Log() de la fonction de survie (FSC).

Log(-Log(FSC)) : activez cette option si vous souhaitez que XLSTAT affiche le Log(–Log()) de la fonction de survie (FSC).

Résidus : activez cette option si vous souhaitez que XLSTAT affiche le graphique des résidus en fonction du temps.

Résultats

XLSTAT propose un nombre important de tableaux et de graphiques afin de faciliter l'analyse et l'interprétation des résultats.

Statistiques descriptives : le tableau de statistiques descriptives présente pour toutes les variables sélectionnées des statistiques simples. Pour les variables quantitatives, sont affichés le nombre d'observations, le nombre de données manquantes, le nombre de données non manquantes, la moyenne, et l'écart-type (non biaisé). Pour les variables qualitatives sont affichées les modalités leurs effectifs et pourcentage respectifs.

Synthèse de la sélection des variables : dans le cas où une méthode de sélection a été choisie, XLSTAT affiche la synthèse de la sélection. Dans le cas d'une sélection pas à pas, les statistiques correspondant aux différentes étapes sont affichées.

Coefficients d'ajustement : dans ce tableau est affichée une série de statistiques pour le modèle indépendant (correspondant au cas où il n'y aurait aucune variables dans le modèle) et pour le modèle ajusté.

- **Observations** : le nombre total d'observations prises en compte;
- **DDL** : degrés de liberté;
- **-2 Log(Vrais.)** : le logarithme de la fonction de vraisemblance associée au modèle;
- **AIC** : le critère d'information d'Akaike (Akaike's Information Criterion);
- **SBC** : le critère bayésien de Schwarz (Schwarz's Bayesian Criterion);
- **Itérations** : nombre d'itérations nécessaires à la convergence de l'algorithme.

Test de l'hypothèse nulle H0 : **beta=0** : l'hypothèse H0 correspond au modèle indépendant (sans variables explicatives) ; on cherche à vérifier si le modèle ajusté est significativement plus performant que ce modèle. Trois tests sont proposés : le test du rapport des vraisemblance (-2 Log(Vrais.)), le test du Score, et le test de Wald. Les trois statistiques suivent une loi du Khi² dont les degrés de liberté sont indiqués.

Paramètres du modèle : pour chaque paramètre du modèle sont affichés l'estimation du paramètre, l'écart-type correspondant, le Khi² de Wald, la p-value correspondante, ainsi qu'un intervalle de confiance associé.

Les **résidus** sont donnés pour chaque observation.

Les **quantiles** sont soit donnés pour chaque observation et chaque valeur des quantiles dans le cas du modèle de régression, soit pour chaque valeur des quantiles pour les courbes de survie paramétrique.

Les **graphiques** obtenus dépendent des options activées. On peut représenter la fonction de survie cumulée, -log de cette fonction, le log de cette dernière ou la fonction de risque ainsi que les graphiques des résidus.

Exemple

Un exemple de régression de survie paramétrique est disponible sur le Centre d'aide XLSTAT :

http://www.xlstat.com/demo-survregf.htm

Un exemple de courbe de survie paramétrique est disponible sur le Centre d'aide XLSTAT :

http://www.xlstat.com/demo-survcurvef.htm

Bibliographie

Collett D. (1994). Modeling Survival Data In Medical Research. Chapman and Hall, London.

Cox D. R. and Oakes D. (1984). Analysis of Survival Data. Chapman and Hall, London.

Hill C., Com-Nougué C., Kramar A., Moreau T., O'Quigley J. Senoussi R. and Chastang C. (1996). Analyse statistique des données de survie. 2-nd Edition, INSERM, Médecine-Sciences, Flammarion.

Kalbfleisch J. D. and Prentice R. L. (2002). The Statistical Analysis of Failure Time Data. 2-nd edition, John Wiley & Sons, New York.

Appariement des coefficients de propension

Utilisez l'appariement des coefficients de propension pour apparier des participants de deux groupes distincts afin de contrôler l'effet de variables confondantes.

Dans cette section :

Description

Le coefficient de propension est défini comme la probabilité pour un participant d'appartenir à un groupe en connaissant certaines variables dites confondantes. L'appariement des coefficients de propension est une technique qui s'attache à réduire le biais potentiel associé à ces variables confondantes dans les études observationnelles.

Présentation

Un exemple typique dans lequel des variables confondantes peuvent être rencontrées serait une étude visant à évaluer l'effet d'un nouveau médicament. Les participants appartiendraient au groupe traité s'ils ont reçu le nouveau médicament et au groupe contrôle dans le cas contraire. Après un certain temps, l'étude pourrait mesurer le taux de survie pour chacun des groupes. Si le taux de survie pour le groupe traité est plus faible que celui du groupe contrôle, on serait tenté de conclure que le médicament n'apporte aucun bénéfice ou, pire, qu'il semble avoir des effets dangereux sur la santé des participants.

En fait, les deux groups ne sont pas identiques et le nouveau médicament a été administré à un groupe de personnes qui avaient déjà une maladie grave diagnostiquée au commencement de l'étude. Au contraire, le groupe contrôle était constitué d'un groupe de personne relativement en bonne santé où seulement une faible proportion avait une maladie grave déclarée.

Dans cet exemple, la variable "maladie grave détectée" est un factor confondant car sa valeur influe sur la probabilité pour un participant d'appartenir à un groupe plutôt qu'à l'autre. Contrôler les effets de cette variable confondante est fortement recommandé car, sans cela, elle pourrait introduire un biais important dans les résultats expérimentaux. Dans notre cas, les participants du groupe ayant reçu le traitement avaient une maladie sérieuse détectée ce qui pourrait expliquer un taux de mortalité plus élevé dans ce groupe comparé au groupe contrôle.

Les variables confondantes sont inhérentes aux études dans lesquelles la procédure d'affectation d'un participant à un groupe donné n'est pas aléatoire. Les raisons possibles pour

ne pas utiliser de procédures aléatoires sont nombreuses, elles peuvent être d'origine éthique, légale, économique ou tout simplement pratique. L'étude est alors définie comme étant une étude observationnelle ou un essai non-randomisé.

L'appariement des coefficients de propension est l'une des meilleures techniques visant à réduire l'effet de facteurs confondants. L'idée fondamentale est d'estimer la probabilité d'appartenance à un groupe en ajustant un modèle de régression logistique sur la variable groupe avec les facteurs confondants comme prédicteurs. Cette probabilité, appelée coefficient de propension, doit refléter les effets des facteurs confondants identifiés.

Ces coefficients de propension sont ensuite utilisés pour apparier chaque participant du groupe traité au participant le plus similaire issu du groupe contrôle. Finalement, l'ensemble des participants ainsi appairés constitue deux groupes similaires du point de vue des variables confondantes et les biais potentiels dans les résultats expérimentaux devraient s'en trouver réduits.

Estimer les coefficients de propension

La notion de coefficient de propension a été proposée la première fois dans Rosenbaum P.R., Rubin D.B. (1983a) comme l'affectation de traitement conditionnel aux covariables :

$$p_i = \Pr(Z_i = 1|X_i)$$

Avec p_i le coefficient de propension, Z_i l'indicateur de groupe (traité ou contrôle) et X_i faisant référence aux covariables suspectées d'être des facteurs confondants.

Dans XLSTAT, le coefficient de propension est estimé en utilisant une régression logistique ou modèle logit (voir également la Régression Logistique pour une description plus détaillée).

La variable groupe est la variable dépendante du modèle logit. C'est une variable binaire qui sépare les participants appartenant au groupe traité de ceux appartenant au groupe contrôle. Vous pouvez choisir la catégorie indiquant le groupe traité.

Les variables explicatives du modèle logit sont les variables confondantes du modèle de coefficient de propension. Elles peuvent être continues (quantitatives) ou catégorielles (qualitative) et XLSTAT vous permet éventuellement d'estimer les interactions entre ces variables.

Algorithme d'appariement

Une fois que les coefficients de propension ont été estimés, chaque participant du groupe traitement est appairé au participant le plus similaire du groupe contrôle (Rosenbaum P. R. (1989)**)**. La similarité est évaluée comme la distance entre les fonctions logit des coefficients de propension définies par :

$$\text{logit}(p_i) = \log(\frac{p_i}{1-p_i})$$

La matrice de distance est alors calculée entre le groupe traité et le groupe contrôle. L'implémentation d'XLSTAT propose deux métriques pour cela :

- la distance euclidienne ;

- la distance de Mahalanobis.

L'utilisateur peut également mettre une limite haute au-dessus de laquelle la distance entre deux participants est considérée comme trop grande pour autoriser un appariement. On parle de caliper ou rayon de caliper pour faire référence à cette limite et elle est définie comme suit :

$$C = a\sqrt{\frac{\sigma_T^2 + \sigma_C^2}{2}}$$

Avec C le rayon de caliper, σ_T^2 la variance de $\text{logit}(p_i)$ pour le groupe traité, σ_C^2 la variance de $\text{logit}(p_i)$ pour le groupe contrôle et α un coefficient. Il n'y a pas de fort consensus dans la littérature sur la valeur que devrait avoir α. Les valeurs rencontrées les plus fréquemment sont 0.1, 0.2 et 0.25. Ces valeurs et d'autres, moins communes (0.5 et 1) sont proposées dans XLSTAT. L'utilisateur a également la possibilité d'affecter à C la valeur de son choix.

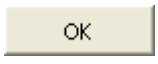
Deux algorithmes sont disponibles dans XLSTAT pour réaliser l'opération d'appariement : l'algorithme glouton et l'algorithme optimal. Avec les deux algorithmes, il est possible d'apparier chaque participant du groupe traité avec un participant du groupe contrôle ou avec un nombre spécifique de participants du groupe contrôle ou bien encore avec tous les participants du groupe contrôle. Il est à noter que les deux dernières configurations peuvent donner des résultats significativement différents selon l'algorithme d'appariement choisi.

L'algorithme glouton procède en appariant successivement chaque participant du groupe traité avec le meilleur candidat parmi les candidats disponibles dans le groupe contrôle. Par disponible, il est entendu que les candidats doivent être à une distance inférieure au rayon de caliper (si l'option caliper est activée) et que l'algorithme glouton réalise les appariements sans remise. Ainsi, une fois qu'un participant du groupe contrôle a été appairé à un participant du groupe traité, il n'est plus disponible pour les prochaines opérations d'appariement. L'ordre initial des participants du groupe traité a donc un effet sur le résultat final de l'opération d'appariement. Pour modérer cette caractéristique de l'algorithme glouton, XLSTAT propose une randomisation aléatoire des participants qui peut être activée.

L'algorithme optimal de son côté ne souffre pas de cette limitation. Il est basé sur une optimisation du flot de coût minimum qui minimise la distance totale pour tous les participants à l'opération d'appariement. L'appariement reste donc sans remise mais la notion d'ordre n'est plus un problème. Avec cet algorithme, une option supplémentaire est disponible pour équilibrer l'appariement en fonction d'une variable qualitative de groupe. Cette option devrait être utilisée lorsqu'il est attendu que la fréquence d'occurrence de chaque catégorie de cette variable dans le groupe de traitement soit reproduite dans le groupe contrôle appairé.

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas (mode colonnes), XLSTAT considère que les sites sont en lignes et les objets/variables en colonnes. Si la flèche est vers la droite (mode lignes), XLSTAT considère que les objets/variables sont en lignes et les sites en colonnes.

Onglet **Général** :

Variable de groupe : sélectionnez la variable de groupe que vous souhaitez modéliser. Elle doit être une variable binaire indiquant si un participant appartient au groupe qui a reçu le traitement ou bien le groupe de contrôle. Si des en-têtes de colonnes ont été sélectionnés, veuillez vérifier que l'option « Libellés des variables » est activée.

Modalité traitement : sélectionnez la modalité de la variable de groupe qui correspond au groupe qui a reçu le traitement. Les modalités devraient être détectées automatiquement et proposées dans le menu déroulant lorsque la variable de groupe est sélectionnée. Si la détection ne fonctionne pas ou bien si la liste déroulante ne correspond plus à la variable après un changement de sélection, vous pouvez cliquer sur le bouton juste à droite du menu déroulant pour rafraichir l'affichage.

Variables explicatives :

Quantitatives : activez cette option si vous voulez inclure une ou plusieurs variables explicatives quantitatives dans le modèle. Sélectionnez alors la ou les variables correspondantes sur la feuille Excel. Les données sélectionnées doivent être de type numérique. Si le libellé des variables a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

Qualitatives : activez cette option si vous voulez inclure une ou plusieurs variables explicatives qualitatives dans le modèle. Sélectionnez alors la ou les variables correspondantes sur la feuille Excel. Les données sélectionnées peuvent être de tout type, mais les données numériques sont automatiquement considérées comme nominales. Si le libellé des variables a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée (voir *description*).

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des variables : activez cette option si la première ligne des données sélectionnées (variables dépendantes et explicatives, libellés des observations) contient un libellé.

Libellés des observations : activez cette option si vous voulez utiliser des libellés d'observations pour l'affichage des résultats. Si l'option « Libellés des variables » est activée, la première cellule de la sélection doit comprendre un en-tête. Si vous n'activez pas cette option, des libellés seront automatiquement créés (Obs1, Obs2, …).

Onglet **Options** :

Tolérance : entrez la valeur de la tolérance seuil en deçà de laquelle une variable est automatiquement ignorée.

Intervalle de confiance (%) : entrez l'étendue en pourcentage de l'intervalle de confiance à utiliser pour les différents tests, et pour le calcul des intervalles de confiance autour des paramètres et des prédictions. Valeur par défaut : 95.

Interactions / Niveau : activez cette option pour inclure des interactions dans le modèle puis entrez le niveau maximum d'interaction (valeur comprise entre 1 et 4).

Méthode de Firth : activez cette option pour utiliser la vraisemblance pénalisée de Firth (voir description). Cette méthode n'est pas disponible pour les modèles logit multinomial et ordinal.

Conditions d'arrêt :

- **Itérations** : entrez le nombre maximal d'itérations pour l'algorithme de Newton-Raphson. Les calculs sont interrompus dès que le nombre maximal d'itérations est dépassé. Valeur par défaut : 100.
- **Convergence** : entrez la valeur seuil d'évolution maximale de log vraisemblance d'une itération à l'autre, qui une fois atteinte permet de considérer que l'algorithme a convergé. Valeur par défaut : 0,000001.

Mélanger les lignes : activez cette option pour mélanger les lignes (participants) avant de lancer l'algorithme glouton d'appariement. Cette option est disponible uniquement lorsque l'algorithme glouton a été sélectionné.

Equilibrage des groupes : activez cette option si vous voulez utiliser une variable catégorielle pour équilibrer les groupes entre le groupe traité et le groupe contrôle lors de l'appariement. Ensuite sélectionnez la variable correspondante dans le feuillet Excel. Les données sélectionnées peuvent être de n'importe quel type mais les données numériques seront automatiquement considérées comme catégorielle. Si un en-tête de colonne a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

Méthode d'appariement :

Algorithme glouton : activez cette option si vous voulez utiliser l'algorithme glouton durant l'appariement.

Algorithme optimal : activez cette option si vous voulez utiliser l'algorithme optimal durant l'appariement.

Distance euclidienne / de Mahalanobis : sélectionnez le type de distance que vous voulez utiliser pour réaliser l'opération d'appariement.

Nombre d'appariement :

Une à une : activez cette option pour appairer chaque participant du groupe traité à un participant du groupe contrôle à la condition qu'il y ait des candidats qui conviennent.

Une à plusieurs : activez cette option pour appairer chaque participant du groupe traité avec un nombre déterminé de participant du groupe contrôle à la condition qu'il y ait des candidats qui conviennent. Ensuite, saisissez le nombre de participants voulu.

Une à toutes : activez cette option pour appairer tous les participants du groupe contrôle à un participant du groupe traité.

Caliper : activez cette option pour fixer une limite sur la distance acceptable pour un appariement. Ensuite sélectionnez la valeur du rayon de caliper que vous voulez utiliser dans le menu déroulant. Si vous sélectionnez « Définie par l'utilisateur », vous devez également saisir cette valeur.

Onglet **Données manquantes** :

Ne pas accepter les valeurs manquantes : activez cette option pour que XLSTAT empêche la poursuite des calculs si des valeurs manquantes sont détectées.

Supprimer les observations : activez cette option pour supprimer les observations comportant des données manquantes.

Estimer les données manquantes : activez cette option pour estimer les données manquantes avant le début des calculs.

- **Moyenne ou mode** : activez cette option pour estimer les données manquantes en utilisant la moyenne (variables quantitatives) ou le mode (variables qualitatives) pour les variables correspondantes.
- **Plus proche voisin** : activez cette option pour estimer les données manquantes d'une observation en recherchant le plus proche voisin de l'observation.

Onglet **Sorties** :

Statistiques descriptives : activez cette option pour afficher les statistiques descriptives pour les variables sélectionnées.

Coefficients d'ajustement : activez cette option pour afficher le tableau des statistiques d'ajustement du modèle.

Test de Hosmer-Lemeshow : activez cette option pour afficher les résultats du test de Hosmer-Lemeshow.

Analyse de type II : activez cette option pour afficher le tableau d'analyse de la variable de type II.

Coefficients normalisés : activez cette option pour afficher les paramètres normalisés du modèle (coefficients bêta).

Résumé de l'appariement : activez cette option pour afficher un résumé des observations appairées.

Coefficient de propension : activez cette option si vous voulez afficher la liste de tous les coefficients de propension.

Matrice de distance : activez cette option si vous voulez afficher la matrice de distance.

Détail des observations appairées : activez cette option si vous voulez afficher les observations appairées.

Onglet Graphiques :

Coefficients normalisés : activez cette option pour afficher les paramètres standardisés du modèle avec leurs intervalles de confiance sur un graphique.

Courbe ROC : activez cette option pour afficher la courbe ROC.

Box plot des coefficients : activez cette option si vous voulez afficher le box plot des coefficients de propension pour chaque groupe.

Fonction de survie cumulée : activez cette option pour afficher les graphiques relatifs à la fonction de survie cumulée.

-Log(FSC) : activez cette option si vous souhaitez que XLSTAT affiche le –Log() de la fonction de survie (FSC).

Log(-Log(FSC)) : activez cette option si vous souhaitez que XLSTAT affiche le Log(–Log()) de la fonction de survie (FSC).

Résidus : activez cette option si vous souhaitez que XLSTAT affiche le graphique des résidus en fonction du temps.

Résultats

XLSTAT propose un nombre important de tableaux et de graphiques afin de faciliter l'analyse et l'interprétation des résultats.

Statistiques descriptives : le tableau de statistiques descriptives présente pour toutes les variables sélectionnées des statistiques simples. Pour les variables quantitatives Sont affichés le nombre d'observations, le nombre de données manquantes, le nombre de données non manquantes, la moyenne, et l'écart-type (non biaisé). Pour les variables qualitatives, dont la variable dépendante, sont affichées les modalités leurs effectifs et pourcentage respectifs.

Coefficients d'ajustement : dans ce tableau est affichée une série de statistiques pour le modèle indépendant (correspondant au cas où la combinaison linéaire des variables explicatives se réduit à une constante) et pour le modèle ajusté.

- **Observations** : le nombre total d'observations prises en compte (somme des poids des observations) ;
- **Somme des poids** : le nombre total d'observations prises en compte (somme des poids des observations multipliés par les poids dans la régression) ;
- **DDL** : degrés de liberté ;
- **-2 Log(Vrais.)** : le logarithme de la fonction de vraisemblance associée au modèle;
- **R² (McFadden)** : coefficient compris comme le R² entre 0 et 1 qui mesure le bon ajustement du modèle. Ce coefficient est égal à 1 moins le rapport de la vraisemblance du modèle ajusté sur la vraisemblance du modèle indépendant ;
- **R²(Cox et Snell)** : coefficient compris comme le R² entre 0 et 1 qui mesure le bon ajustement du modèle. Ce coefficient est égal à 1 moins le rapport de la vraisemblance du modèle ajusté sur la vraisemblance du modèle indépendant, le rapport étant porté à l'exposant 2/Sw, où Sw est la somme des poids ;
- **R²(Nagelkerke)** : coefficient compris comme le R² entre 0 et 1 qui mesure le bon ajustement du modèle. Ce coefficient est égal au rapport du R² de Cox et Snell, divisé par 1 moins le la vraisemblance du modèle indépendant portée à l'exposant 2/Sw ;
- **AIC** : le critère d'information d'Akaike (Akaike's Information Criterion) ;
- **SBC** : le critère bayésien de Schwarz (Schwarz's Bayesian Criterion).

Test de l'hypothèse nulle H0 : **Y=p0** : l'hypothèse H0 correspond au modèle indépendant qui donne la probabilité p0 quelques soient les valeurs des variables explicatives ; on cherche à vérifier si le modèle ajusté est significativement plus performant que ce modèle. Trois tests sont proposés : le test du rapport des vraisemblance (-2 Log(Vrais.)), le test du Score, et le test de Wald. Les trois statistiques suivent une loi du Khi² dont les degrés de liberté sont indiqués.

Analyse de Type II : ce tableau n'a d'intérêt que s'il y a plus d'une variable explicative. On test ici le modèle ajusté contre un test dont on aurait retiré la variable de la ligne du tableau en question. Si la probabilité Pr > LR est inférieur à un seul de signification que l'on se fixe (typiquement 0.05), alors la contribution de la variable à l'ajustement du modèle est significative. Sinon, elle peut être retirée du modèle.

Le tableau des **coefficients normalisés** (aussi appelés coefficients bêta) permet de comparer le poids relatif des variables. Plus la valeur absolue d'un coefficient est élevée, plus le poids de la variable correspondante est important. Lorsque l'intervalle de confiance autour des coefficients normalisés comprend la valeur 0 (cela est facilement visible sur le graphique des coefficients normalisés), le poids d'une variable dans le modèle n'est pas significatif.

Courbe ROC : la courbe ROC permet d'évaluer la performance du modèle au travers de l'aire sous la courbe (AUC) et de comparer plusieurs modèles entre eux (voir la section description pour plus de détails).

Le tableau de **résumé des appariements** affiche des indicateurs sur les proportions de participants qui ont été appairés. Le coût total en termes de distance est également donné juste en dessous du tableau.

Le tableau des **coefficients de propension** donne les coefficients de propension calculés pour chaque participant des deux groups. La valeur du logit de ces coefficients est également donnée. C'est cette dernière valeur qui sera utilisé pour calculer la distance séparant chaque participant. Les bornes inférieures et supérieures sont également données pour ces deux variables.

La **matrice de distance** permet d'avoir une vue générale sur toutes les distances. Les participants appartenant au groupe traité sont sur les lignes alors que les participants du groupe contrôle sont sur les colonnes. Les distances correspondant à des paires sont affichées en gras.

Le **box plot** affiche plusieurs paramètres de la distribution du logit des coefficients de propension pour le groupe traité au complet, la sous partie du groupe traité appairée, la sous partie du groupe contrôle appairée et le groupe contrôle en entier.

Exemple

Un exemple d'appariement des coefficients de propension est disponible sur le Centre d'aide XLSTAT :

http://www.xlstat.com/customer/fr/portal/articles/2826861http://www.xlstat.com/customer/fr/portal/articles/282

Bibliographie

Rosenbaum P.R., Rubin D.B. (1983a). The central role of the propensity score in observational studies for causal effects. *Biometrika* ;70:41–55

Rosenbaum P. R. (1989). Optimal matching for observational studies. *Journal of the American Statistical Association* ;Vol. 84, No. 408, 1024-1032

Sensibilité et Spécificité

Utilisez cet outil pour calculer entre autres les valeurs de sensibilité, spécificité, odds ratio, les valeurs prédictives, et les rapports de vraisemblance associés à un test de dépistage ou à une méthode de détection. Ces indices permettent d'évaluer en médecine la performance d'un test utilisé pour diagnostiquer une maladie ou, en contrôle qualité la présence d'un défaut dans un produit issu d'une chaîne de fabrication.

Dans cette section :

Description

Cette méthode a d'abord été développée pendant la seconde guerre mondiale pour la mise au point de moyens efficaces de détection des avions japonais. Elle a ensuite été appliquée de manière plus générale en détection du signal, puis en médecine, où elle est aujourd'hui très utilisée.

La problématique est la suivante : on étudie un phénomène, souvent de nature binaire (par exemple, la présence ou absence d'une maladie) et on souhaite mettre au point un test permettant de détecter efficacement la survenance d'un événement précis (par exemple, la présence de la maladie).

Soit une V variable binaire ou multinomiale décrivant le phénomène pour N individus suivis. Notons par + les individus pour lesquels l'événement se produit, et par – ceux pour lesquels il ne se produit pas. Soit T un test dont le but est de détecter si l'événement se produit ou non. T peut être une variable binaire (présence/absence), qualitative (par exemple la couleur), ou quantitative (par exemple une concentration). Pour les variables binaires ou qualitatives, soit t1 la modalité de la variable correspondant à la survenance de l'événement étudié. Pour une variable quantitative, t1 est une valeur seuil en-deçà ou au-delà de laquelle l'événement se produit.

Une fois le test appliqué à l'ensemble des N individus, on obtient un tableau individus/variables dans lequel, pour chaque individu, est consignée la survenance ou non de l'événement, ainsi que le résultat du test.

A	B	C
Individu	Maladie	Test
I1	+	+
I2	+	+
I3	+	+
I4	+	+
I5	+	+
I6	+	+
I7	-	-
I8	+	-
I9	-	-
I10	-	-
I11	-	-

L	M	N
Individu	**Maladie**	**T1**
I1	+	0
I2	+	0,1
I3	+	0,2
I4	+	0,3
I5	-	0,4
I6	+	0,5
I7	-	1
I8	-	2
I9	-	3
I10	-	4

Cas d'un test binaire Cas d'un test quantitatif

Ces tableaux sont parfois transformés en un tableau de contingence 2 x 2, plus synthétique :

	M+	**M-**
T+	25	12
T-	8	13

Dans l'exemple ci-dessus, on a 25 individus pour lesquels le test a bien détecté la présence de la maladie et 13 pour lesquels il a bien détecté l'absence de maladie. En revanche, pour 20 individus le diagnostique est mauvais puisque, pour 8 d'entre eux, le test conclut à l'absence de la maladie alors que le patient est malade, et pour 12 d'entre eux, il conclut qu'ils sont malades, alors qu'ils ne le sont pas.

On utilise le vocabulaire suivant :

Vrais positifs (VP) : nombre d'individus déclarés positifs par le test et qui le sont effectivement.

Faux positifs (FP) : nombre d'individus déclarés positifs par le test mais qui sont en réalité négatifs.

Vrais négatifs (VN) : nombre d'individus déclarés négatifs par le test et qui le sont effectivement.

Faux négatifs (FN) : nombre d'individus détectés négatifs par le test mais qui sont en réalité positifs.

Plusieurs indices synthétiques ont été mis au point afin d'évaluer la performance d'un test :

Sensibilité (aussi appelée **Fraction de Vrais Positifs**): proportion d'individus positifs effectivement bien détectés par le test. Autrement dit, la sensibilité permet de mesurer à quel point le test est performant lorsqu'il est utilisé sur des individus positifs. Le test est parfait pour les individus positifs lorsque la sensibilité vaut 1, équivalent à un tirage au hasard lorsque la sensibilité vaut 0.5. S'il est inférieur à 0.5, le test est contre- performant et on aurait intérêt à inverser la règle pour qu'il soit supérieur à 0.5 (à condition que cela n'affecte pas la spécificité). La définition mathématique est : Sensibilité = VP/(VP + FN).

Spécificité (aussi appelée **Fraction de Vrais Négatifs**): proportion d'individus négatifs effectivement bien détectés par le test. Autrement dit, la spécificité permet de mesurer à quel point le test est performant lorsqu'il est utilisé sur des individus négatifs. Le test est parfait pour

les individus négatifs lorsque la spécificité vaut 1, équivalent à un tirage au hasard lorsque la spécificité vaut 0.5. S'il est inférieur à 0.5, le test est contre- performant et on aurait intérêt à inverser la règle pour qu'il soit supérieur à 0.5 (à condition que cela n'affecte pas la sensibilité). La définition mathématique est : Spécificité = VN/(VN + FP).

Fraction de faux positifs (FFP) : proportion de négatifs détectés comme des positifs par le test (1-Spécificité).

Fraction de faux négatifs (FFN) : proportion de positifs détectés comme des négatifs par le test (1-Sensibilité)

Prévalence de l'événement : fréquence de survenance de l'événement dans l'échantillon total (VP+FN)/N.

Valeur Prédictive Positive : proportion de cas effectivement positifs parmi les positifs détectés par le test. On a VPP = VP/(VP+FP), ou VPP=Sensibilité x Prévalence / [(Sensibilité x Prévalence + (1-Spécificité)(1- Prévalence)]. C'est une valeur fondamentale qui a la particularité de dépendre aussi de la prévalence, une donnée indépendante de la qualité du test.

Valeur Prédictive Négative : proportion de cas effectivement négatifs parmi les négatifs détectés par le test. On a VPN = VN/(VN+FN), ou VPP= Spécificité(1- Prévalence) / [Spécificité(1- Prévalence) + (1- Sensibilité)Prévalence]. Cet indice dépend aussi de la prévalence, une donnée indépendante de la qualité du test.

Rapport de vraisemblance positif (LR+) : ce rapport indique à quel point un individu a plus de chance d'être positif en réalité si le test est positif. On a LR+ = Sensibilité / (1-Spécificité).

Rapport de vraisemblance négatif (LR-) : ce rapport indique à quel point un individu a plus de chance d'être positif en réalité, si le test est négatif. Le risque relatif est nécessairement une valeur positive ou nulle. On a LR- = (1-Sensibilité) / (Spécificité).

Odds ratio : l'odds ratio indique à quel point un individu a plus de chance d'être positif si le test est positif, par rapport au cas où le test est négatif. Par exemple, un odds ratio de 2 signifie que la chance pour que l'événement se produise est 2 fois supérieure si le test est positif. L'odds ratio est une valeur positive ou nulle. On a Odds ratio = VPxVN /(FPxFN).

Risque relatif : le risque relatif est un ratio qui mesure à quel point le test se comporte mieux lorsqu'il est positif par rapport au cas où il est négatif. Par exemple, un risque relatif de 2 signifie que le test est 2 fois plus performant lorsqu'il est positif que lorsqu'il est négatif. Une valeur proche de 1 correspond à un cas d'indépendance entre les lignes et les colonnes, et à un test aussi performant quand il est positif que lorsqu'il est négatif. Le risque relatif est une valeur positive ou nulle donnée par : Risque relatif = VP/(VP+FP) / (FN/(FN+VN)).

Intervalles de confiance

Pour les différents indices synthétiques présentés ci-dessus, plusieurs méthodes de calcul de leur variance et donc des intervalles de confiance on été proposées. On distingue deux

familles : la première concerne les proportions, comme par exemple la sensibilité ou la spécificité, et la seconde, les ratios tels LR+, LR- l'odds ratio et le risque relatif.

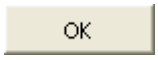
Pour les proportions, XLSTAT propose notamment les estimateurs de Wald simple (Wald, 1939) ou ajusté (Agresti et Coull, 1998), le calcul basé sur le score de Wilson (Wilson, 1927), avec éventuellement une correction de continuité, ou l'intervalle de Clopper-Pearson (1934). Agresti et Caffo recommandent d'utiliser plutôt l'intervalle de Wald ajusté ou le score de Wilson.

Pour les ratios, les variances sont calculées suivant une seule méthode, avec ou sans correction de continuité.

Une fois cette variance des indices calculée on suppose la normalité asymptotique de la statistique (ou de son logarithme pour les ratios) pour déterminer l'intervalle de confiance. Hormis les ratios, les différents indices présentés ci-dessus sont des proportions comprises entre 0 et 1. Si les intervalles sortent de ces limites, XLSTAT corrige automatiquement les bornes de l'intervalle.

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas (mode colonnes), XLSTAT considère que les sites sont en lignes et les objets/variables en colonnes. Si la flèche est vers la droite (mode lignes), XLSTAT considère que les objets/variables sont en lignes et les sites en colonnes.

Onglet **Général**:

Format des données :

Tableau 2x2 (Test\Evénement) : choisissez cette option si vos données sont consignées dans un tableau de contingence 2x2 avec en ligne les tests et en colonne les événements positifs et négatifs. Vous pouvez ensuite préciser dans quelle colonne du tableau se trouvent les événements positifs, et sur quelle ligne se trouvent les cas ou le test de détection est positif.

L'option « Libellés inclus » doit être activée si les libellés des lignes et des colonnes ont été sélectionnés.

Données individuelles : choisissez cette option si vos données sont consignées dans un tableau individus/variables. Vous devez alors sélectionner les **données événement** correspondant au phénomène étudié (par exemple la présence ou l'absence maladie) et préciser quel code est associé à l'**événement positif** (par exemple M ou + pour un individu malade). Vous devez aussi sélectionner les **données test** correspondant à la valeur du test de diagnostique. Ce test peut-être quantitatif (une concentration), binaire (positif ou négatif) ou qualitatif (une couleur). Si le test de nature quantitative, vous devez préciser si l'on doit considérer qu'il est positif lorsque la valeur test est supérieure ou inférieure à une valeur donnée. Si le test est de nature binaire ou qualitative, vous devez sélectionner la valeur correspondant à un test positif.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés inclus : activez cette option si la première ligne et la première colonne des données sélectionnées contient un libellé. NB : cette option n'est visible que si vous avez sélectionné un tableau de contingence.

Libellés des variables : activez cette option si, en mode colonnes, la première ligne des données sélectionnées contient un libellé, ou si en mode lignes, la première colonne des données sélectionnées contient un libellé. NB : cette option n'est visible que si vous avez sélectionné un tableau individus/variables.

Poids : activez cette option si vous voulez pondérer les individus. Si vous n'activez pas cette option, les poids seront tous considérés comme valant 1. Les poids doivent être impérativement supérieurs ou égaux à 0. Si un en- tête de colonne a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

Onglet **Options**:

Intervalles de confiance :

- **Taille(%)** : entrer la taille de l'intervalle confiance souhaitée en % (valeur par défaut : 95).
- **Wald** : activez cette option si vous voulez calculer les intervalles de confiance sur les indices en utilisant l'approximation de la loi binomiale par la loi normale. Activez l'option « Ajusté » pour appliquer l'ajustement d'Agresti et Coull.

- **Wilson score** : activez cette option si vous voulez calculer les intervalles de confiance sur les indices en utilisant l'approximation du Wilson score.
- **Clopper-Pearson** : activez cette option si vous voulez calculer les intervalles de confiance sur les indices en utilisant l'approximation de Clopper-Pearson.
- **Correction de continuité** : activez cette option si vous voulez appliquer la correction de continuité au Wilson score et aux intervalles sur les ratios.

Prévalence a priori : si vous savez que la maladie concerne une proportion donnée d'individus dans la population totale, vous pouvez utiliser cette information pour ajuster les valeurs prédictives calculées à partir de votre échantillon.

Onglet **Données manquantes** :

Ne pas accepter les valeurs manquantes : activez cette option pour que XLSTAT empêche la poursuite des calculs si des valeurs manquantes sont détectées.

Supprimer les observations : activez cette option pour supprimer les observations comportant des données manquantes.

Résultats

Les résultats sont constitués du tableau de contingence, suivi du tableau des différents indices décrits dans la section description.

Exemple

Un exemple de calcul des indices de sensibilité et spécificité est disponible sur le Centre d'aide XLSTAT :

http://www.xlstat.com/demo-sensf.htm

Bibliographie

Agresti A. (1990). Categorical Data Analysis. John Wiley and Sons, New York.

Agresti A., and Coull B.A. (1998). Approximate is better than "exact" for interval estimation of binomial proportions. *The American Statistician*, **52**, 119-126.

Agresti A. and Caffo, B. (2000). Simple and effective confidence intervals for proportions and differences of proportions result from adding two successes and two failures. *The American Statistician*, **54**, 280-288.

Clopper C.J. and Pearson E.S. (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26, 404-413.

Newcombe R. G. (1998). Two-sided confidence intervals for the single proportion: comparison of seven methods. *Statistics in Medicine*, **17**, 857-872.

Zhou X.H., Obuchowski N.A., McClish D.K. (2002). Statistical Methods in Diagnostic Medicine. John Wiley & Sons.

Pepe M.S. (2003). The Statistical Evaluation of Medical Tests for Classification and Prediction, Oxford University Press.

Wilson, E.B. (1927). Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, **22**, 209-212.

Wald, A., & Wolfowitz, J. (1939). Confidence limits for continuous distribution functions. The Annals of Mathematical Statistics, 10, 105-118.

Courbes ROC

Utilisez cet outil pour générer une courbe ROC, permettant de représenter l'évolution de la proportion de vrais positifs (aussi appelée sensibilité) en fonction de la proportion de faux positifs (correspondant à 1 moins la spécificité), et d'évaluer un critère de classification binaire tel un test de dépistage de maladie, ou un contrôle de la présence de défauts sur un produit sorti d'une chaîne de fabrication.

Dans cette section :

Description

Les courbes ROC (de l'anglais *Receiver Operating Characteristic*) ont d'abord été développées pendant la seconde guerre mondiale pour la mise au point de moyens efficaces de détection des avions japonais. Cette méthodologie a ensuite été appliquée de manière plus générale en détection du signal, puis en médecine, où elle est aujourd'hui très utilisée.

La problématique est la suivante : on étudie un phénomène, souvent de nature binaire (par exemple, la présence ou l'absence d'une maladie) et on souhaite mettre au point un test permettant de détecter efficacement la survenance d'un événement précis (par exemple, la présence de la maladie).

Soit une V variable binaire ou multinomiale décrivant le phénomène pour N individus suivis. Notons par + les individus pour lesquels l'événement se produit, et par – ceux pour lesquels il ne se produit pas. Soit T un test dont le but est de détecter si l'événement se produit ou non. T est une le plus souvent quantitative (par exemple une concentration), mais elle peut être qualitative ordinale (représentant des niveaux).

On souhaite définir la valeur seuil en-deçà ou au-delà de laquelle l'événement se produit. On étudie alors un ensemble de valeurs seuil possibles et, pour chacune, on calcule différentes statistiques dont les plus simples sont :

- **Vrais positifs (VP)** : nombre d'individus déclarés positifs par le test et qui le sont effectivement.
- **Faux positifs (FP)** : nombre d'individus déclarés positifs par le test mais qui sont en réalité négatifs.
- **Vrais négatifs (VN)** : nombre d'individus déclarés négatifs par le test et qui le sont effectivement.

- **Faux négatifs (FN)** : nombre d'individus détectés négatifs par le test mais qui sont en réalité positifs.
- **Prévalence de l'événement** : fréquence de survenance de l'événement dans l'échantillon total (VP+FN)/N.

Plusieurs indices synthétiques ont été mis au point afin d'évaluer la performance du test à une valeur seuil donnée :

Sensibilité (aussi appelée **Fraction de Vrais Positifs**): proportion d'individus positifs effectivement bien détectés par le test. Autrement dit, la sensibilité permet de mesurer à quel point le test est performant lorsqu'il est utilisé sur des individus positifs. Le test est parfait pour les individus positifs lorsque la sensibilité vaut 1, équivalent à un tirage au hasard lorsque la sensibilité vaut 0.5. S'il est inférieur à 0.5, le test est contre- performant et on aurait intérêt à inverser la règle pour qu'il soit supérieur à 0.5 (à condition que cela n'affecte pas la spécificité). La définition mathématique est : Sensibilité = VP/(VP + FN).

Spécificité (aussi appelée **Fraction de Vrais Négatifs**): proportion d'individus négatifs effectivement bien détectés par le test. Autrement dit, la spécificité permet de mesurer à quel point le test est performant lorsqu'il est utilisé sur des individus négatifs. Le test est parfait pour les individus négatifs lorsque la spécificité vaut 1, équivalent à un tirage au hasard lorsque la spécificité vaut 0.5. S'il est inférieur à 0.5, le test est contre- performant et on aurait intérêt à inverser la règle pour qu'il soit supérieur à 0.5 (à condition que cela n'affecte pas la sensibilité). La définition mathématique est : Spécificité = VN/(VN + FP).

Fraction de faux positifs (FFP) : proportion de négatifs détectés comme des positifs par le test (1-Spécificité).

Fraction de faux négatifs (FFN) : proportion de positifs détectés comme des négatifs par le test (1-Sensibilité)

Valeur Prédictive Positive : proportion de cas effectivement positifs parmi les positifs détectés par le test. On a VPP = VP/(VP+FP), ou VPP=Sensibilité x Prévalence / [(Sensibilité x Prévalence + (1-Spécificité)(1- Prévalence)]. C'est une valeur fondamentale qui a la particularité de dépendre aussi de la prévalence, une donnée indépendante de la qualité du test.

Valeur Prédictive Négative : proportion de cas effectivement négatifs parmi les négatifs détectés par le test. On a VPN = VN/(VN+FN), ou VPP= Spécificité(1- Prévalence) / [Spécificité(1- Prévalence) + (1- Sensibilité)Prévalence]. Cet indice dépend aussi de la prévalence, une donnée indépendante de la qualité du test.

Rapport de vraisemblance positif (LR+) : ce rapport indique à quel point un individu a plus de chance d'être positif en réalité, si le test est positif. On a LR+ = Sensibilité / (1-Spécificité).

Rapport de vraisemblance négatif (LR-) : ce rapport indique à quel point un individu a plus de chance d'être positif en réalité, si le test est négatif. Le risque relatif est nécessairement une valeur positive ou nulle. On a LR- = (1-Sensibilité) / (Spécificité).

Exactitude : l'exactitude est le rapport (VP+VN)/(VP+VN+FP+FN). Plus elle proche de 1 meilleur est le test.

Intervalles de confiance

Pour les différents indices synthétiques présentés ci-dessus, plusieurs méthodes de calcul de leur variance et donc des intervalles de confiance on été proposées. On distingue deux familles : la première concerne les proportions, comme par exemple la sensibilité ou la spécificité, et la seconde, les ratios tels LR+ et LR-.

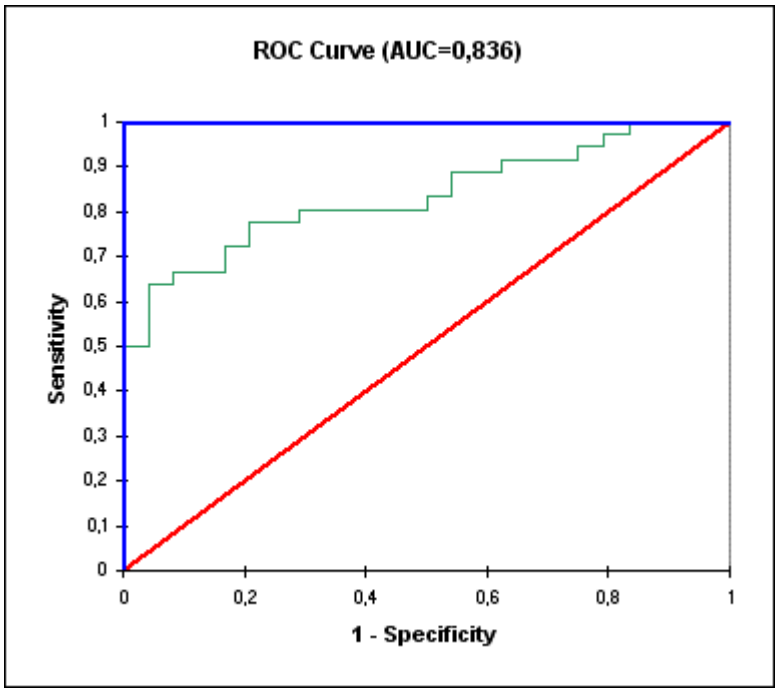
Pour les proportions, XLSTAT propose notamment les estimateurs de Wald simple (Wald, 1939) ou ajusté (Agresti et Coull, 1998), le calcul basé sur le score de Wilson (Wilson, 1927), avec éventuellement une correction de continuité, ou l'intervalle de Clopper-Pearson (1934). Agresti et Caffo recommandent d'utiliser plutôt l'intervalle de Wald ajusté ou le score de Wilson.

Pour les ratios, les variances sont calculées suivant une seule méthode, avec ou sans correction de continuité.

Une fois cette variance des indices calculée on suppose la normalité asymptotique de la statistique (ou de son logarithme pour les ratios) pour déterminer l'intervalle de confiance. Hormis les ratios, les différents indices présentés ci-dessus sont des proportions comprises entre 0 et 1. Si les intervalles sortent de ces limites, XLSTAT corrige automatiquement les bornes de l'intervalle.

Courbe ROC

La courbe ROC correspond à la représentation graphique du couple (1 – spécificité ; sensibilité) pour les différentes valeurs seuil. Son allure est en soit en escalier (comme ci-dessous) soit en droites par morceaux.



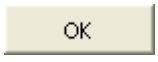
L'aire sous la courbe (ou *Area Under the Curve – AUC*) est un indice synthétique calculé pour les courbes ROC. L'AUC correspond à la probabilité pour qu'un événement positif soit classé comme positif par le test sur l'étendue des valeurs seuil possibles. Pour un modèle idéal, on a AUC=1 (ci- dessus en bleu), pour un modèle aléatoire, on a AUC=0.5 (ci-dessus en rouge). On considère habituellement que le modèle est bon dès lors que la valeur de l'AUC est supérieure à 0.7. Un modèle bien discriminant doit avoir une AUC entre 0.87 et 0.9. Un modèle ayant une AUC supérieure à 0.9 est excellent.

Sen (1960), Bamber (1975) et Hanley et McNeil (1982) ont proposé différentes méthodes de calcul de la variance de l'AUC. Toutes sont proposées par XLSTAT. XLSTAT propose un test de comparaison de l'AUC à la valeur de 0.5 qui correspond à un test aléatoire. Ce test s'appuie sur la différence entre l'AUC et 0.5 divisée par la variance calculée selon l'une des trois méthodes proposées. La statistique obtenue est supposée suivre une loi normale standard, ce qui permet notamment le calcul de la p-value.

L'AUC peut aussi être utilisée pour comparer différents tests entre eux. Si les différents tests ont été appliqués à différents groupes d'individus, les échantillons sont indépendants. Dans ce cas, XLSTAT utilise un test de Student pour comparer les AUC (ce qui suppose la normalité de l'AUC, ce qui est acceptable si les échantillons ne sont pas trop petits). Si différents tests ont été appliqués aux mêmes individus, on est dans un cas de données appariées. Dans ce cas, XLSTAT calcule la matrice de covariance des AUC sur la base des travaux de Sen (1960), rapportés par Delong et Delong (1988), pour pouvoir ensuite calculer la variance de la différence entre deux AUC, puis pour calculer la p-value associée sur la base d'une hypothèse de normalité.

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas (mode colonnes), XLSTAT considère que les sites sont en lignes et les objets/variables en colonnes. Si la flèche est vers la droite (mode lignes), XLSTAT considère que les objets/variables sont en lignes et les sites en colonnes.

Onglet **Général** :

Données évènement : sélectionnez les données correspondant au phénomène étudié (par exemple la présence ou l'absence de maladie) et précisez quel code est associé à l'**évènement positif** (par exemple M ou + pour un individu malade).

Données test : sélectionnez les données correspondant à la valeur du test de diagnostique. Les données doivent être quantitatives. S'il s'agit de données ordinales, elles doivent être recodées en données quantitatives (par exemple 0,1,2,3,4). Vous devez préciser si l'on doit considérer qu'il est positif lorsque la valeur test est supérieure ou inférieure à une valeur seuil déterminée au cours des calculs.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des variables : activez cette option si, en mode colonnes, la première ligne des données sélectionnées contient un libellé, ou si en mode lignes, la première colonne des données sélectionnées contient un libellé.

Poids : activez cette option si vous voulez pondérer les individus. Si vous n'activez pas cette option, les poids seront tous considérés comme valant 1. Les poids doivent être impérativement supérieurs ou égaux à 0. Si un en-tête de colonne a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

Onglet **Options** :

Intervalles de confiance :

- **Taille (%)** : entrez la taille de l'intervalle confiance souhaitée en % (valeur par défaut : 95).
- **Wald** : activez cette option si vous voulez calculer les intervalles de confiance sur les indices en utilisant l'approximation de la loi binomiale par la loi normale. Activez l'option « Ajusté » pour appliquer l'ajustement d'Agresti et Coull.
- **Wilson score** : activez cette option si vous voulez calculer les intervalles de confiance sur les indices en utilisant l'approximation du Wilson score.
- **Clopper-Pearson** : activez cette option si vous voulez calculer les intervalles de confiance sur les indices en utilisant l'approximation de Clopper-Pearson.

- **Correction de continuité** : activez cette option si vous voulez appliquer la correction de continuité au Wilson score et aux intervalles sur les ratios.

Prévalence a priori : si vous savez que la maladie concerne une proportion donnée d'individus dans la population totale, vous pouvez utiliser cette information pour ajuster les valeurs prédictives calculées à partir de votre échantillon.

Test pour l'AUC : vous avez la possibilité de comparer l'AUC (aire sous la courbe) à 0.5, la valeur qu'elle aurait si la variable test de classification était purement aléatoire. Ce test est réalisé en utilisant la méthode de calcul de la variance choisie ci-dessus.

Coûts : activez cette option si vous voulez évaluer le coût associé aux différentes décisions possibles fonction des valeurs de la variable test. Pour cela entrez les coûts associés aux différentes situations : VP (vrais positifs), FP (faux positifs), FN (faux négatifs), VN (vrais négatifs).

Onglet **Prétraitement** :

Données manquantes :

Ne pas accepter les valeurs manquantes : activez cette option pour que XLSTAT empêche la poursuite des calculs si des valeurs manquantes sont détectées.

Supprimer les observations : activez cette option pour supprimer les observations comportant des données manquantes.

Ignorer les données manquantes : activez cette option pour ignorer les données manquantes.

Groupes :

Analyse par groupe : activez cette option puis sélectionnez ici les données d'appartenance à des groupes si vous souhaitez que les calculs soient effectués sur chaque groupe séparément.

- **Comparer** : activez cette option si vous souhaitez que les courbes soient comparées pour les différents groupes, et si vous souhaitez que les tests de comparaison soient calculés.

Filtrer : activez cette option puis sélectionnez ici les données d'appartenance à des groupes, si vous souhaitez que les calculs ne soient effectués que sur certains groupes. Une boîte de dialogue apparaîtra au début des calculs, pour vous permettre de choisir les groupes actifs. Si l'option « Analyse par groupe » est activée, l'analyse ne sera effectuée groupe par groupe, que pour les groupes sélectionnés.

Onglet **Sorties** :

Statistiques descriptives : activez cette option pour afficher les statistiques descriptives pour les variables sélectionnées.

Analyse ROC : activez cette option pour afficher le tableau des différents indices calculés pour chaque valeur de la variable test. Vous pouvez choisir d'afficher ou de ne pas afficher les valeurs prédictives, les rapports de vraisemblance et les vrais/faux positifs et négatifs.

Test pour l'AUC : activez cette option si souhaitez que soient affichés les résultats de comparaison de l'AUC à la valeur 0.5 correspondant à une règle de classification aléatoire.

Comparaison des AUC : si vous avez sélectionné plusieurs variables de test ou une variable de groupe, activez cette option pour comparer les AUC obtenues pour les différentes variables ou les différents groupes.

Onglet **Graphiques** :

Courbe ROC : activez cette option pour afficher la courbe ROC.

Vrais/Faux +/- : activez cette option pour afficher le graphique en barres empilées présentant les % de VP/VN/FP/FN pour les différentes valeurs de la variable test. L'option est disponible uniquement si l'option Vrais/Faux +/- dans l'onglet sorties est activée.

Graphique de décision : activez cette option pour afficher le graphique de décision de votre choix. Ce graphique doit vous aider à choisir quel est le bon niveau de la variable test à retenir pour une performance optimale du test.

Comparaison des courbes ROC : activez cette option pour afficher sur un même graphique les courbes ROC associées aux différentes variables tests ou aux différents groupes. Cette option n'est disponible que si vous avez sélectionné plusieurs variables ou si une variable de groupe a été sélectionnée.

Résultats

Statistiques descriptives : dans un premier tableau vous trouverez les statistiques pour le ou les tests sélectionnés, suivi d'un tableau rappelant pour le phénomène étudié le nombre d'occurrences de chaque événement et la prévalence de l'événement positif dans l'échantillon. La ligne en gras correspond à l'événement positif.

Courbe ROC : La courbe ROC est ensuite affichée. La ligne en pointillée allant de (0 ;0) à (1 ;1) correspond à la courbe d'un test aléatoire. La ligne colorée et continue correspond à la courbe ROC. Les petits carrés correspondent aux données observées (il y a un carré par valeur observée de la variable test).

Analyse ROC : ce tableau présente pour chaque valeur seuil possible de la variable test les différents indices présentés dans la section description. Sous le tableau vous trouverez le rappel de la règle fixée dans la boîte de dialogue pour identifier les positifs par rapport à la valeur seuil. Sous le tableau est affiché un graphique en barres empilées permettant de visualiser l'évolution des VP, VN, FP, FN en fonction de la valeur seuil. Si l'option correspondante a été activée, le **graphique de décision** est ensuite affiché (par exemple évolution du coût en fonction de la valeur seuil).

Aire sous la courbe (AUC) : ce tableau donne la valeur de l'AUC, la variance et un intervalle de confiance.

Comparaison de l'AUC à 0.5 : ces résultats permettent de comparer le test choisi à un test qui serait simplement une règle de classification aléatoire. L'intervalle de confiance correspond à celui de la différence. Les différentes statistiques sont ensuite présentées et notamment la p-value, suivi de l'interprétation du test de comparaison.

Comparaison des AUC : si vous avez sélectionné plusieurs variables de test, une fois les résultats ci-dessus affichés pour chaque variable, vous trouverez la matrice de covariance des AUC, suivi du tableau des différences pour chaque couple avec en commentaire l'intervalle de confiance, puis du tableau des p-value. Les valeurs en gras correspondent aux différences significatives. Un graphique permettant de comparer les **courbes ROC** est ensuite affiché.

Exemple

Un exemple de calcul de courbes ROC est disponible sur le Centre d'aide XLSTAT :

http://www.xlstat.com/demo-rocf.htm

Un exemple de calcul et de comparaison de courbes ROC est disponible sur le Centre d'aide XLSTAT :

http://www.xlstat.com/demo-roccomparef.htm

Bibliographie

Agresti A. (1990). Categorical Data Analysis. John Wiley and Sons, New York.

Agresti A., and Coull B.A. (1998). Approximate is better than "exact" for interval estimation of binomial proportions. *The American Statistician*, **52**, 119-126.

Agresti A. and Caffo, B. (2000). Simple and effective confidence intervals for proportions and differences of proportions result from adding two successes and two failures. *The American Statistician*, **54**, 280-288.

Bamber D. (1975). The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of Mathematical Psychology*, **12**, 387-415.

Clopper C.J. and Pearson E.S. (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26, 404-413.

DeLong E.R., DeLong D.M., Clarke-Pearson D.L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, **44(3)**, 837-845.

Hanley J.A. and McNeil B.J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, **143**, 29-36.

Hanley J. A. and McNeil B. J. (1983). A method of comparing the area under two ROC curves derived from the same cases. *Radiology*, **148**, 839-843.

Newcombe R. G. (1998). Two-sided confidence intervals for the single proportion: comparison of seven methods. *Statistics in Medicine*, **17**, 857-872.

Pepe M.S. (2003). The Statistical Evaluation of Medical Tests for Classification and Prediction, Oxford University Press.

Sen P. K. (1960). On some convergence properties of U-statistics. *Calcutta Statistical Association Bulletin*, **10**, 1-18.

Wilson, E.B. (1927). Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, **22**, 209-212.

Wald, A., & Wolfowitz, J. (1939). Confidence limits for continuous distribution functions. The Annals of Mathematical Statistics, 10, 105-118.

Zhou X.H., Obuchowski N.A., McClish D.K. (2002). Statistical Methods in Diagnostic Medicine. John Wiley & Sons.

Modèle Illness-Death paramétrique

Utilisez le modèle Illness-Death afin de déterminer le temps de survie et de modéliser les différents états parcourus au cours du temps. Des variables explicatives peuvent compléter le modèle pour analyser l'effet d'une ou plusieurs variables sur les transitions entre états.

Ce modèle s'intègre dans le cadre des méthodes pour l'analyse de survie.

Dans cette section :

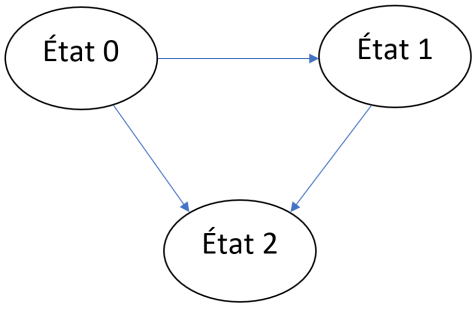
Description

Lorsque l'on dénombre plus de 2 états en survie, on parle alors de modèle multi-états. Le modèle Illness-Death est un modèle multi-états particulier qui fait intervenir 3 états : l'état initial, l'état transitoire et l'état absorbant notés respectivement 0, 1, 2.

Ce modèle est largement utilisé en épidémiologie pour observer l'évolution d'une maladie, l'influence d'une maladie sur la mortalité ou encore de la mortalité post-opératoire. On retrouve aussi de nombreuses applications dans le domaine actuariel pour calculer les coûts d'assurances et des contrats viatiques.

Modèle

Dans XLSTAT, nous ne considérons que les modèles Illness-Death irréversibles représentés ci-dessous :



Soit X le processus stochastique résumant l'évolution des états des individus au cours du temps. Deux fonctions caractérisent X : les **intensités de transition** et les **probabilités de transition**. Dans le modèle Illness-Death, on estime les intensités de transition qui sont l'analogue de la fonction de risque en analyse de survie.

On suppose que le processus X est markovien non-homogène. Les probabilités et intensités de transition associées à la transition $k \to l$ avec k et l des états sont définies telles que :

- Les probabilités de transition p_{kl} ne dépendent que de l'état au temps présent s

$$p_{kl}(s,t) = \mathbb{P}(X(t) = l | X(s) = k)$$

- Les intensités de transition $\alpha_{kl}(t)$ varient en fonction du temps.

$$\alpha_{kl}(t) = \lim_{\Delta t \to 0} \frac{p_{kl}(t, t+\Delta t)}{\Delta t}$$

L'estimation des intensités de transition par maximum de vraisemblance tient compte des éventuelles censures et troncatures.

Censures et troncatures

En analyse de survie les censures et troncatures représentent un manque d'information lors du recueil des données. Si on les ignore, les estimations seront faussées. Le modèle Illness-Death de XLSTAT accepte les censures à droite et par intervalle ainsi que les troncatures à gauche sous l'hypothèse que les censures et troncatures sont non-informatives.

Les censures interviennent lorsque l'on ne connait pas la date exacte de survenue de l'événement tandis que les troncatures interviennent lorsque l'événement ne peut pas se produire avant ou après un seuil donné.

- La censure à droite se produit lorsque le sujet n'a pas subi l'événement absorbant à la fin de sa période d'observation. En d'autres mots, si on considère t le temps de la survenue de l'événement et t_{fin} le temps de fin d'observation alors $t > t_{fin}$
- La censure par intervalle se produit lorsque le sujet subit l'événement entre deux observations. En d'autres mots, si on considère t_l et t_{l+1} les temps de deux visites consécutives alors la survenue de l'événement au temps t s'est produite telle que $t \in [t_l,\ t_{l+1}]$
- La troncature à gauche se produit lorsque l'individu ne peut pas subir l'événement avant le début de l'étude.

Estimation des intensités de transition

Dans un **modèle Illness-Death paramétrique**, les intensités de transition de base suivent une loi de Weibull avec a_{kl} et b_{kl} les paramètres de forme et d'échelle.

Pour la transition $k \to l$ on écrit donc l'intensité de transition de base $\alpha_{0,kl}$:

$$ \alpha_{0,kl}(t) = a_{kl} \: \Big(\frac{1}{b_{kl}}\Big)^{a_{kl}} \: t^{a_{kl}-1}. $$

Les formules de survie S et de densité f associées s'écrivent alors : $$ S(t) = e^{ - \Big(\dfrac{t}{b} \Big)^{a}} \qquad ;\qquad f(t) = a\: \Big(\dfrac{1}{b} \Big)^{a} \: t^{a-1} \: e^{ - \Big(\dfrac{t}{b} \Big)^{a}}. $$

Les paramètres a_{kl} et b_{kl} sont estimés par maximum de vraisemblance.

Ajout de variables explicatives

Dans un modèle Illness-Death, les intensités de transition permettent de prendre en compte l'effet de variables explicatives. Par exemple l'effet d'un traitement sur une maladie ou sur la mortalité.

D'après le modèle à risques proportionnels de Cox, on obtient le modèle à intensités proportionnelles tel que pour la transition $k \to l$:

$$\alpha_{kl}(t) = \alpha_{0,kl}(t)\, e^{\beta_{kl}^T Z_{kl}}.$$

L'intensité de transition $\alpha_{kl}(t)$ s'exprime en fonction de l'intensité de base $\alpha_{0,kl}$ (estimée par la loi de Weibull), des vecteurs des effets β_{kl} et de la matrice des variables explicatives Z_{kl}.

Ce modèle induit 2 hypothèses qui sont à vérifier a posteriori :

- L'hypothèse de proportionnalité suppose que le rapport des intensités est constant :

$$\frac{\alpha(t|Z_i)}{\alpha(t|Z_j)} = \frac{\alpha_0(t) \times e^{\beta^T Z_i}}{\alpha_0(t) \times e^{\beta^T Z_j}} = e^{\beta^T (Z_i - Z_j)}.$$

Il n'existe pas de méthode simple pour vérifier cette hypothèse. Il faut donc estimer, sur chaque intervalle de temps, les intensités de transition puis vérifier que leur rapport est constant quel que soit t.

Pour éviter cette méthode fastidieuse, on peut considérer l'âge comme échelle de temps plutôt que comme variable explicative. Cette méthode ne garantit pas la proportionnalité, mais la favorise (voir Touraine, 2013).

- L'hypothèse de log-linéarité qui suppose que les intensités de transition à l'échelle log sont linéaires :

$$log(\alpha(t|Z)) = log(\alpha_0(t)) + \beta_1 Z_1 + ... + \beta_p Z_p.$$

Il n'existe pas de méthode simple pour vérifier cette deuxième hypothèse. Cependant l'ajout de variables à 2 modalités permet de contourner l'hypothèse. Il est donc conseillé à l'utilisateur d'inclure uniquement des variables binaires en tant que variables explicatives ou des variables quantitatives avec peu d'écart.

La formule du modèle à intensités proportionnelles permet de calculer les intensités de transition par covariables dans XLSTAT en fixant la variable Z_{kl} à la valeur des modalités associées.

Les β_{kl} sont estimés par maximum de vraisemblance.

Vraisemblance

Dans un modèle Illness-Death irréversible, les probabilités de transition et la vraisemblance s'expriment en fonction des α_{kl} et de la fonction de survie S.

La vraisemblance doit inclure les différents types de censure. On a donc 7 expressions de vraisemblance possibles représentant les différentes combinaisons de censure. Ces expressions sont explicitées dans les travaux de Touraine C. (2013).

Chaque individu n'étant pas exposé au même type de censure, la vraisemblance est calculée pour chaque individu. La vraisemblance totale L est le produit des contributions individuelles L_i :

$$L = \prod_{i=1}^{n} L_i$$

N'ayant pas de solution explicite du maximum de vraisemblance, les paramètres sont estimés d'après l'algorithme itératif de Levenberg-Marquardt.

Algorithme de Levenberg-Marquardt (LM)

L'estimation des intensités de transition du modèle Illness-Death se fait en maximisant la log-vraisemblance avec l'algorithme de Levenberg-Marquardt.

Cette procédure itérative associe deux méthodes d'optimisation :

- La méthode de descente du gradient : maximise la log-vraisemblance et met à jour les paramètres à chaque itération dans la direction du gradient
- La méthode de Newton-Raphson : maximise la dérivée de la log-vraisemblance et met à jour les paramètres à chaque itération.

Lors de l'optimisation, l'algorithme agit comme la méthode de descente du gradient lorsque les paramètres sont loin de la solution et comme la méthode de Newton-Raphson lorsque les paramètres sont proches de la solution. Ainsi, l'algorithme de LM est plus rapide que la méthode du gradient, longue à converger, et plus efficace que la méthode de Newton, coûteuse en temps de calcul.

Prédictions : Probabilités de transition et espérances de vie

Les prédictions permettent à l'utilisateur de donner les probabilités de transition et l'espérance de vie des individus entre 2 temps :

- le temps d'entrée dans l'étude,
- le temps de sortie de l'étude.

Parmi les probabilités de transition, on a :

- p_{00} la probabilité de rester dans l'état initial,
- p_{01} la probabilité de passer de l'état initial à l'état transitoire,
- p_{02} la probabilité de passer de l'état initial à l'état absorbant,
- p_{11} la probabilité de rester dans l'état transitoire,
- p_{12} la probabilité de passer de l'état transitoire à l'état absorbant.

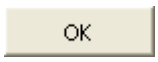
Parmi les espérances de vie, on a :

- E_{00} l'espérance de vie non-malade : c'est-à-dire le temps que l'on peut espérer pour un individu de rester dans l'état initial sachant qu'il y est déjà,
- E_{02} l'espérance de vie au sens commun : c'est-à-dire le temps que l'on peut espérer pour un individu de vivre sachant qu'il est dans l'état initial,
- E_{12} l'espérance de vie d'un individu malade : c'est-à-dire le temps qu'on peut espérer pour un individu de vivre sachant qu'il est dans l'état transitoire.

Les formules explicites de ces probabilités et espérances de vie sont données dans les travaux de Touraine C. (2013).

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas (mode colonnes), XLSTAT considère que les objets/variables sont en colonnes. Si la flèche est vers la droite (mode lignes), XLSTAT considère que les objets/variables sont en lignes.

Onglet **Général**:

Indicateurs d'état : sélectionnez ici les 2 colonnes qui correspondent respectivement à l'état transitoire et à l'état absorbant, les individus étant tous dans l'état initial au début de l'étude. Ces 2 colonnes sont binaires, mettez 0 pour indiquer que l'individu n'a pas été dans l'état et 1 sinon.

Données de date : sélectionnez ici les données de temps qui correspondent aux 4 âges clés : - Âge d'entrée : âge de l'individu lorsqu'il entre dans l'étude - Borne gauche : âge de l'individu - avant la transition dans l'état transitoire si l'individu y est allé, - sinon aux dernières nouvelles. - Borne droite : âge de l'individu - au moment de la transition dans l'état transitoire si l'individu y est allé, - sinon aux dernières nouvelles. - Âge de dernières nouvelles : âge de l'individu - lors de la transition dans l'état absorbant si l'individu y est allé, - sinon à la dernière visite.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule au préalable.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des colonnes : activez cette option si la première ligne de la sélection contient le libellé des variables.

Poids : activez cette option si vous voulez pondérer les observations. Si vous n'activez pas cette option, les poids seront tous considérés comme valant 1. Les poids doivent être des entiers impérativement supérieurs ou égaux à 0. Si un en-tête de colonne a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

Onglet **Covariables** :

Covariables : activez cette option si vous voulez inclure une ou plusieurs variables explicatives au modèle.

Différentes par transition : activez cette option si vous voulez personnaliser vos covariables sur chaque transition. Dans le cas contraire, les variables explicatives seront les mêmes pour toutes les transitions.

Covariables : État 0 -> État 1 si "Différentes par transition" cochée, **Covariables** sinon : - **Quantitatives** : activez cette option si vous voulez inclure une ou plusieurs variables explicatives quantitatives dans le modèle. Sélectionnez alors la ou les variables correspondantes sur la feuille Excel. Les données sélectionnées doivent être de type numérique. Si l'option **Différentes par transition** est activée les covariables sont valables uniquement pour la transition "État 0 -> État 1", si l'option n'est pas cochée les variables sélectionnées sont appliquées à chaque transition.

- **Qualitatives** : activez cette option si vous voulez inclure une ou plusieurs variables explicatives qualitatives dans le modèle. Sélectionnez alors la ou les variables correspondantes sur la feuille Excel. Les données sélectionnées peuvent être de tout type, mais les données numériques sont automatiquement considérées comme nominales. Si l'option **Différentes par transition** est activée les covariables sont valables uniquement pour la transition "État 0 -> État 1", si l'option n'est pas cochée les variables sélectionnées sont appliquées à chaque transition.

Si le libellé des variables a été sélectionné, veuillez vérifier que l'option « Libellés des colonnes » est activée.

Si l'option **"Différentes par transition"** est cochée :

Covariables : État 0 -> État 2" : - **Quantitatives** : activez cette option si vous voulez inclure une ou plusieurs variables explicatives quantitatives dans le modèle pour la transition État 0 -> État 2. Sélectionnez alors la ou les variables correspondantes sur la feuille Excel. Les données sélectionnées doivent être de type numérique. - **Qualitatives** : activez cette option si vous voulez inclure une ou plusieurs variables explicatives qualitatives dans le modèle pour la transition État 0 -> État 2. Sélectionnez alors la ou les variables correspondantes sur la feuille

Excel. Les données sélectionnées peuvent être de tout type, mais les données numériques sont automatiquement considérées comme nominales.

Si le libellé des variables a été sélectionné, veuillez vérifier que l'option « Libellés des colonnes » est activée.

Covariables : État 1 -> État 2" : - **Quantitatives** : activez cette option si vous voulez inclure une ou plusieurs variables explicatives quantitatives dans le modèle pour la transition État 1 -> État 2. Sélectionnez alors la ou les variables correspondantes sur la feuille Excel. Les données sélectionnées doivent être de type numérique. - **Qualitatives** : activez cette option si vous voulez inclure une ou plusieurs variables explicatives qualitatives dans le modèle pour la transition État 1 -> État 2. Sélectionnez alors la ou les variables correspondantes sur la feuille Excel. Les données sélectionnées peuvent être de tout type, mais les données numériques sont automatiquement considérées comme nominales.

Si le libellé des variables a été sélectionné, veuillez vérifier que l'option « Libellés des colonnes » est activée.

Libellés des colonnes : activez cette option si la première ligne des variables sélectionnées contient un libellé.

Onglet **Options**:

Niveau de signification (%) : entrez la valeur du niveau de signification à utiliser pour les tests (valeur par défaut : 5%). Cette valeur est aussi utilisée pour déterminer les intervalles de confiance pour les statistiques calculées.

Les options suivantes sont dédiées aux conditions d'arrêt de l'algorithme LM :

- **Itérations** : entrez le nombre maximal d'itérations pour l'algorithme de Levenberg-Marquardt. Les calculs sont interrompus dès que le nombre maximal d'itérations est dépassé. Valeur par défaut : 100.
- **Convergence** : entrez la valeur seuil d'évolution maximale qui une fois atteinte permet de considérer que l'algorithme a convergé.
    - **Paramètres** : la somme des carrés du gradient (mise à jour des paramètres). Valeur par défaut : 0,00001.
    - **Vraisemblance** : la différence absolue de la log-vraisemblance d'une itération à l'autre. Valeur par défaut : 0,00001.
    - **Dérivées** : du calcul des dérivées et de la hessienne. Valeur par défaut : 0,001.

Onglet **Prédictions** :

Sous-onglet **Général** :

Prédiction : activez cette option si vous souhaitez faire de la prédiction. Si vous activez cette option et que des variables explicatives ont été ajoutées dans l'onglet principal **Covariables** de la fenêtre, vous pouvez choisir d'inclure des variables explicatives à la prédiction grâce au sous-onglet **Covariables** (voir ci-dessous).

Temps d'entrée : entrez l'âge auquel les individus prédits entrent dans l'étude.

Temps de sorties : entrez l'âge auquel les individus prédits sortent de l'étude.

Sous-onglet **Covariables** :

Covariables : activez cette option si vous voulez inclure une ou plusieurs variables explicatives aux prédictions. Si vous choisissez cette option, vous devez veiller à ce que les données de prédiction soient organisées comme les données d'estimation : mêmes variables et même ordre de sélection. Cette option est disponible uniquement si des covariables ont été ajoutées dans le modèle dans l'onglet **Covariables**.

Libellés des variables : activez cette option si la première ligne de la sélection contient le libellé des variables.

Libellés des observations : activez cette option si vous voulez utiliser des libellés des observations disponibles sur une feuille Excel pour l'affichage des résultats. La première ligne ne doit pas comprendre d'en-tête. Si vous n'activez pas cette option, des libellés seront automatiquement créés (PredObs1, PredObs2, …).

Onglet **Données manquantes** :

Ne pas accepter les valeurs manquantes : activez cette option pour que XLSTAT empêche la poursuite des calculs si des valeurs manquantes sont détectées.

Supprimer les observations : activez cette option pour supprimer les observations comportant des données manquantes.

Onglet **Sorties** :

Statistiques descriptives : activez cette option pour afficher les statistiques descriptives des indicateurs d'états et des données de temps.

Log-vraisemblance : activez cette option pour afficher les valeurs de la log-vraisemblance avec et/ou sans covariables.

Paramètres de Weibull : activez cette option pour afficher les paramètres de Weibull calculés pour les 3 transitions.

Paramètres de régression : activez cette option pour afficher le tableau des coefficients du modèle. La première colonne indique la transition et la deuxième colonne les variables associées. La troisième colonne indique la valeur des coefficients. La quatrième colonne indique l'erreur standard du coefficient qui mesure le degré de précision de l'estimation. Les cinquième et sixième colonnes fournissent les résultats du test de significativité de Wald avec respectivement la statistique de Wald et la p-valeur. Les trois dernières colonnes indiquent la valeur du risque relatif (Hazard Ratio) et l'intervalle de confiance associé. Cette option est disponible uniquement si des covariables ont été ajoutées au modèle dans l'onglet "Covariables".

Distribution de survie : activez cette option pour afficher le tableau des probabilités de survie par transition.

- **Intervalle de confiance** : activez cette option pour ajouter les intervalles de confiance associés au tableau des probabilités de survie.
- **Par covariable** : activez cette option pour afficher les probabilités de survie par covariable. Cette option est uniquement disponible si des variables explicatives qualitatives ont été ajoutées au modèle.

Probabilités de transition : activez cette option pour afficher le tableau des probabilités de transition.

- **Intervalle de confiance** : activez cette option pour ajouter les intervalles de confiance associés au tableau des probabilités de transition.
- **Par covariable** : activez cette option pour afficher les probabilités de transition par covariable. Cette option est uniquement disponible si des variables explicatives qualitatives ont été ajoutées au modèle.

Intensités de transition : activez cette option pour afficher le tableau des intensités de transition.

- **Intervalle de confiance** : activez cette option pour ajouter les intervalles de confiance associés au tableau des intensités de transition.
- **Par covariable** : activez cette option pour afficher les intensités de transition par covariable. Cette option est uniquement disponible si des variables explicatives qualitatives ont été ajoutées au modèle.

Onglet **Graphiques** :

Distribution de survie : activez cette option pour afficher le graphique relatif aux fonctions de survie pour chacune des transitions.

- **Intervalle de confiance** : activez cette option pour ajouter les intervalles de confiance associés au graphique des fonctions de survie.
- **Par covariable** : activez cette option pour afficher en plus les fonctions de survie par covariables. Cette option est uniquement disponible si des variables explicatives qualitatives ont été ajoutées au modèle.

Probabilités de transition : activez cette option pour afficher le graphique des probabilités de transition.

- **Intervalle de confiance** : activez cette option pour ajouter les intervalles de confiance associés au graphique des probabilités de transition.
- **Par covariable** : activez cette option pour afficher les probabilités de transition par covariables. Cette option est uniquement disponible si des variables explicatives qualitatives ont été ajoutées au modèle.

Intensités de transition : activez cette option pour afficher le graphique des intensités de transition.

- **Intervalle de confiance** : activez cette option pour ajouter les intervalles de confiance associés au graphique des intensités de transition.
- **Par covariable** : activez cette option pour afficher les intensités de transition par covariables. Cette option est uniquement disponible si des variables explicatives qualitatives ont été ajoutées au modèle.

Résultats

XLSTAT propose un nombre important de tableaux et de graphiques afin de faciliter l'analyse et l'interprétation des résultats.

Statistiques descriptives : le tableau des statistiques descriptives présente pour les indicateurs d'états et les données de temps des statistiques simples. Pour les temps, sont affichés le nombre d'observations, le nombre de données manquantes, le nombre de données non-manquantes, la moyenne, et l'écart-type (non biaisé). Pour les indicateurs d'états, sont affichées les modalités, leurs effectifs et pourcentages respectifs.

Coefficients de régression et **Paramètres de Weibull** : ces tableaux affichent les paramètres de Weibull et de régression. Les paramètres de Weibull permettent de calculer les intensités de base. Si le modèle inclut des variables, alors pour chaque variable sont affichés l'estimation du coefficient de régression, l'écart-type correspondant, le χ^2 de Wald, la p-value correspondante. Par ailleurs, le Hazard Ratio (exponentielle du coefficient) est donné ainsi qu'un intervalle de confiance associé.

Intensités de transition : ce tableau affiche les intensités de transition qui sont les pendants de la fonction de risque d'un modèle de survie. Observer les intensités de transition permet de comparer les risques de transition entre états.

Probabilités de transition : ce tableau affiche les probabilités de transition. Observer les probabilités de transition permet de comparer les transitions entre états. Ces quantités sont plus facilement interprétables et plus intuitives que les intensités de transition.

Fonctions de survie : ce tableau affiche les fonctions de survie. Observer les fonctions de survie permet de voir la probabilité de quitter un état. Ces quantités sont plus facilement interprétables et plus intuitives que les intensités de transition.

Prédictions : ce tableau affiche les probabilités de transition prédites et les 3 espérances de vie : au sens commun, de rester malade (rester dans l'état transitoire) et de rester sain (rester dans l'état initial). Les prédictions sont calculées en fonction du temps d'entrée et du temps de sortie de l'étude indiqués par l'utilisateur. Si des variables explicatives ont été ajoutées lors de la prédiction alors les résultats sont donnés pour chaque individu dont on a rentré les covariables associées, sinon les résultats correspondent à un seul individu.

Exemple

Un exemple d'utilisation du modèle Illness-Death paramétrique est disponible sur le Centre d'aide XLSTAT à l'adresse suivante :

http://www.xlstat.com/demo-msmf.htm

Bibliographie

Bourmouche, L. (2016). Modèles multi-états markoviens en analyse de survie.

Commenges, D., & Gégout-Petit, A. (2007). Likelihood for generally coarsened observations from multistate or counting process models. Scandinavian journal of statistics, 34(2), 432-450.

Hinchliffe, S. R., Scott, D. A., & Lambert, P. C. (2013). Flexible parametric illness-death models. The Stata Journal, 13(4), 759-775.

Joly, P., Commenges, D., Helmer, C., & Letenneur, L. (2002). A penalized likelihood approach for an illness–death model with interval-censored data: application to age-specific incidence of dementia. Biostatistics, 3(3), 433-443.

Putter, H., Fiocco, M., & Geskus, R. B. (2007). Tutorial in biostatistics: competing risks and multi-state models. Statistics in medicine, 26(11), 2389-2430.

Saint-Pierre, P. (2005). Modèles multi-états de type markovien et application à l'asthme (Doctoral dissertation, Université Montpellier I).

Touraine, C. (2013). Modèles illness-death pour données censurées par intervalle: application à l'étude de la démence (Doctoral dissertation, Bordeaux 2).

Touraine, C., Gerds, T. A., & Joly, P. (2017). SmoothHazard: An R package for fitting regression models to interval-censored observations of illness-death models. Journal of Statistical Software, 79, 1-22.

Analyse de données de laboratoires

Comparaison de méthodes

Utilisez cet outil pour comparer une nouvelle méthode à une méthode de référence ou, simplement à une méthode disponible. Des tests, des intervalles de confiance et des graphiques tels que le graphique de Bland et Altman sont utilisés pour évaluer les performances de la méthode étudiée. Avec cet outil, vous êtes en mesure de répondre aux recommandations du Clinical and Laboratory Standards Institute (CLSI).

Dans cette section :

Description

Lors de la mise au point d'une nouvelle méthode pour mesurer la concentration ou la quantité d'un élément (molécule, microorganisme, ...), vous pouvez vouloir vérifier si elle donne des résultats similaires à une méthode de référence ou à méthode comparable. S'il y a une différence, il peut être intéressant de savoir si cela est dû à un biais qui dépend ou non de l'endroit où l'on se trouve sur l'échelle de variation. Si une nouvelle méthode de mesure est moins chère qu'une méthode standard et que vous savez qu'il existe un biais, il est possible de prendre en compte ce dernier pour corriger les résultats obtenus.

XLSTAT offre une série d'outils pour évaluer la performance d'une méthode par rapport à une autre.

Analyse de répétabilité

L'analyse de répétabilité et de reproductibilité d'un système de mesure est disponible dans le module XLSTAT-SPC. L'analyse de répétabilité fournie ici est une version allégée qui permet d'analyser la répétabilité de chaque méthode séparément et de les comparer. Pour évaluer la répétabilité d'une méthode, il faut disposer de plusieurs répétitions pour chaque mesure. Les répétitions peuvent être spécifiées en utilisant les "groupes" de la boîte de dialogue (les répétitions doivent avoir le même identifiant). Cela correspond au cas où plusieurs mesures

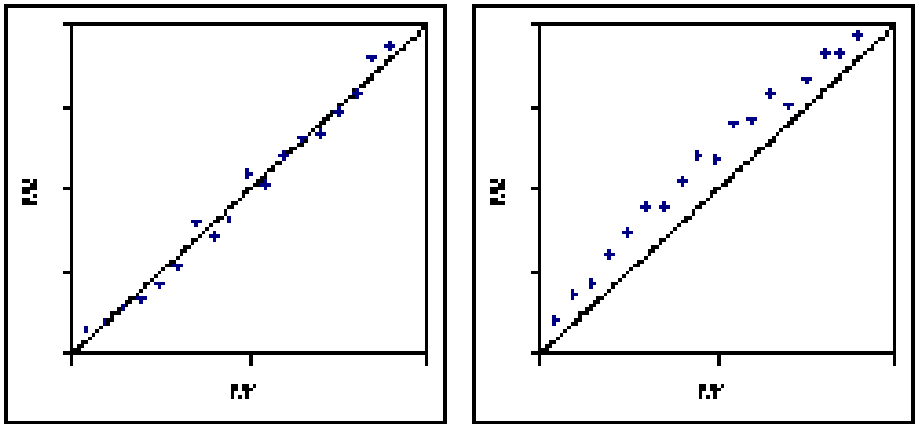
sont effectuées sur un échantillon donné. Si la méthode est répétable, la variance au sein des répétitions doit être faible. XLSTAT calcule la répétabilité comme la racine carrée d'une variance et fournit aussi un intervalle de confiance. Idéalement, l'intervalle de confiance devrait contenir 0.

Test t de Student sur données appariées

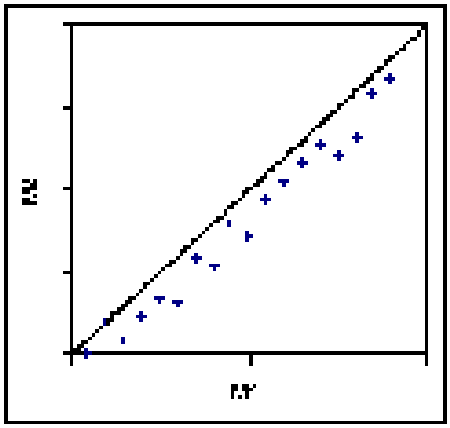
XLSTAT propose parmi les méthodes de comparaison, le test t de Student sur données appariées. Ce test permet de tester l'hypothèse H0 que la moyenne des différences entre les deux méthodes n'est pas différente de 0, contre l'hypothèse alternative Ha qu'elle l'est.

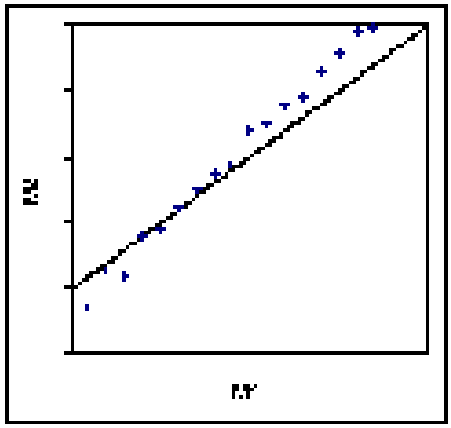
Nuages de points

Dans un premier temps, XLSTAT affiche un nuage de points pour comparer la nouvelle méthode à une méthode de référence. Si les données sont réparties des deux côtés de la bissectrice (qui correspond à une identité parfaite des méthodes) tout en en étant proche, les deux méthodes donnent des résultats cohérents et proches. Si les données sont au-dessus de la ligne, la nouvelle méthode surestime la quantité mesurée. Si les données sont sous la ligne, la nouvelle méthode sous-estime la quantité mesurée, au moins par rapport à la méthode avec laquelle la comparaison est effectuée. Si les données croisent la bissectrice, le biais dépend de là où l'on se trouve sur l'échelle de variation. Si les données sont dispersées de façon aléatoire autour de la bissectrice avec des observations loin d'elle, la nouvelle méthode n'est pas performante.

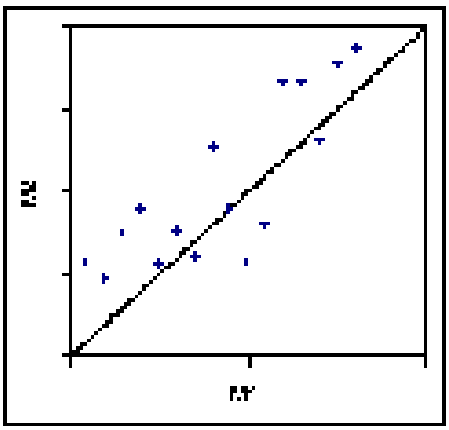


1. Méthodes cohérentes 2. Biais positif constant





3. Biais négatif constant 4. Biais linéaire



5. Méthodes incohérentes

Biais

Le biais est estimé par la moyenne des différences entre les deux méthodes. Si des répétitions sont disponibles, dans un premier temps on calcule la moyenne des répétitions. L'écart-type du biais est calculé, ainsi qu'un intervalle de confiance. Idéalement, cet intervalle de confiance doit comprendre la valeur 0.

Remarque : Le biais est calculé pour le critère qui a été choisi pour l'analyse de Bland Altman (différence, différence en % ou ratio).

Analyse de Bland Altmanet méthodes de comparaison liées

Bland et Altman recommandent de représenter la différence (T-S) entre la méthode en cours d'évaluation (T) et une méthode de référence ou comparable (S) en fonction de la moyenne (T+S)/2 des résultats obtenus pour les deux méthodes. Dans le cas idéal, il ne devrait y avoir aucune corrélation en la différence et la moyenne, qu'il y ait un biais ou non. XLSTAT teste si la corrélation est significativement différente de 0 ou non. Plusieurs possibilités sont proposées pour les ordonnées du graphique : vous pouvez choisir la différence (T-S), la différence en % de la somme 100*(T-S)/(T+S), et le ratio (T/S). Sur le graphique de Bland Altman, XLSTAT affiche

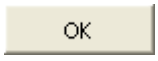
la ligne correspondent au biais, les intervalles de confiance autour du biais, et des différences (ou de la différence en % ou du ratio).

Histogramme et box plot

L'histogramme et le box plot des différences (ou différences en % ou ratio) sont affichés pour valider l'hypothèse que cette quantité est distribuée suivant une loi normale, sachant que cette hypothèse est utilisée pour calculer les intervalles de confiance autour du biais et des différences individuelles. Lorsque l'échantillon est de petite taille, l'histogramme est de peu d'intérêt et il faut seulement considérer le box plot. Si la distribution ne semble pas normale, on peut vérifier ce point avec un test de normalité, et les intervalles de confiance doivent être considérés avec prudence.

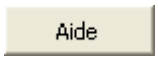
Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.



: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas (mode colonnes), XLSTAT considère que les sites sont en lignes et les objets/variables en colonnes. Si la flèche est vers la droite (mode lignes), XLSTAT considère que les objets/variables sont en lignes et les sites en colonnes.

Onglet **Général**:

Données (M éthod e 1) : sélectionnez les données qui correspondent à la première méthode, ou à la méthode de référence. Si le libellé de la méthode est présent en première position, veillez à ce que l'option « Libellé des variables » soit bien activée.

Données (Méthode 2) : sélectionnez les données qui correspondent à la première méthode, ou à la méthode de référence. Si le libellé de la méthode est présent en première position, veillez à ce que l'option « Libellé des variables » soit bien activée.

Groupes : si des répétitions sont disponibles, activez cette option pour sélectionner les données qui correspondent à l'identifiant du groupe auquel appartient chaque observation. Les observations ayant le même identifiant sont considérées comme des répétitions. XLSTAT calcule les moyennes des répétitions et fournit des résultats sur la répétabilité.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des variables : activez cette option si, en mode colonnes, la première ligne des données sélectionnées contient un libellé, ou si en mode lignes, la première colonne des données sélectionnées contient un libellé.

Onglet **Options**:

Analyse de Bland Altman : activez cette option pour effectuer une analyse de Bland Altman analysis et/ou afficher le graphique de Bland Altman. Vous devez ensuite préciser quelle variable doit être utilisée pour les ordonnées.

Analyse des diffé rences : activez cette option pour effectuer une analyse des différences et/ou afficher le graphique des différences. Vous devez ensuite préciser quelles variables doivent être utilisées pour les abscisses et les ordonnées.

Niveau de signification (%) : entrez le niveau de signification à utiliser pour le test de Student sur données appariées (valeur par défaut : 5 %).

Intervalle de confiance (%) : entrez la taille des intervalles de confiance (valeur par défaut : 95 %).

Onglet **Données manquantes** :

Ne pas accepter les valeurs manquantes : activez cette option pour que XLSTAT empêche la poursuite des calculs si des valeurs manquantes sont détectées.

Supprimer les observations : activez cette option pour supprimer les observations comportant des données manquantes.

Ignorer les données manquantes : activez cette option pour ignorer les données manquantes. Cette option n'est disponible que dans le cas de la présence de groupes.

Onglet **Sorties**:

Statistiques descriptives : activez cette option pour afficher les statistiques descriptives pour les deux méthodes.

Test t sur données appariées : activez cette option pour afficher les résultats correspondant au test t de Student sur données appariées et savoir si les méthodes peuvent être ou non considérées comme non significativement différentes.

Analyse de Bland Altman : activez cette option pour calculer la statistique du biais et les afficher les données utilisées pour la graphique de Bland et Altman.

Onglet **Graphiques** :

Scatter plot : activez cette option pour afficher le nuage de points représentant en abscisse la méthode de référence (ou de comparaison) et en ordonnées la méthode testée.

Graphique de Bland et Altman : activez cette option pour afficher le graphique de Bland et Altman.

Histogramme : activez cette option pour afficher l'histogramme des différences (ou différences en % ou ratios).

Box plot : activez cette option pour afficher le box plot des différences (ou différences en % ou ratios).

Difference plot : activez cette option pour afficher le *difference plot* .

Résultats

Statistiques descriptives : dans un premier tableau vous trouverez les statistiques pour les deux méthodes étudiées.

Test t pour deux échantillons appariés : ces résultats correspondent au test de l'hypothèse H0 que la moyenne des différences entre les deux méthodes n'est pas différente de 0, contre l'hypothèse alternative Ha qu'elle l'est. Une aide à l'interprétation et conclusion sont fournis.

Un graphique en **nuage de points** est affiché pour permettre la comparaison visuelle des deux méthodes. La première bissectrice est affichée sur le graphique. Elle correspond au cas idéal où les échantillons sur lesquels les deux méthodes sont appliquées sont identiques et où les deux méthodes donnent les mêmes résultats.

L'**analyse de Bland Altman** commence par une estimation du biais, en utilisant le critère qui a été choisi (la différence, différence en %, ou le ratio), de l'écart-type de l'intervalle de confiance correspondant. Le graphique de Bland Altman est ensuite affiché pour permettre de visualiser la différence entre les deux méthodes.

L'**histogramme et le box plot** permettent de visualiser comment la différence (ou la différence en % ou le ratio) est distribuée. Une hypothèse de normalité est utilisée, notamment pour le test de Student et pour le calcul des intervalles de confiance. Il convient donc de la valider au moins graphiquement.

Le **différence plot** montre la différence entre les deux méthodes en fonction de la moyenne des deux méthodes ou de la méthode de référence. La ligne correspondant au biais, calculé en utilisant le critère qui a été choisi (la différence, différence en %, ou le ratio), l'écart-type et un intervalle de confiance étant ainsi affichée.

Exemple

Un exemple de comparaison de méthodes est disponible sur le Centre d'aide XLSTAT :

http://www.xlstat.com/demo-blandf.htm

Bibliographie

Altman D.G. and Bland J.M. (1987). Measurement in Medicine: the Analysis of Method Comparison Studies. *The Statistician*, **32**, 307-317.

Bland J.M. and Altman D.G. (2008). Measurement agreement in method comparison studies. *Statistical Methods in Medical Research*; **8**, 135-160.

Hyltoft Petersen P., Stöckl D., Blaabjerg O., Pedersen B., Birkemose E., Thienpont L., Flensted Lassen1 J. and Kjeldsen J. (1997). Graphical interpretation of analytical data from comparison of a field method with a Reference Method by use of difference plots. *Clinical Chemistry*, **43(11)**, 2039-2046.

Bland J. M. and Altman D. G. (2007). Agreement between methods of measurement with multiple observations per individual. *Journal of Biopharmaceutical Statistics*, **17**, 571-582.

Régression de Passing et Bablok

Utilisez cet outil pour comparer deux méthodes de mesure en faisant un minimum d'hypothèses quant à leur distribution.

Dans cette section :

Description

Passing et Bablok (1983) ont mis au point une méthode de régression qui permet de comparer deux méthodes de mesure (par exemple deux techniques de mesure de la concentration d'un analyte) qui s'affranchit des hypothèses lourdes et ici inappropriées de la régression linéaire simple classique. Pour rappel ces hypothèses sont :

- la variable explicative, X dans le modèle $y(i) = a + b \times x(i) + \varepsilon(i)$, est déterministe (pas d'erreurs de mesure),
- la variable Y suit une loi normale de moyenne aX,
- la variance de l'erreur est constante.

Par ailleurs, les valeurs extrêmes peuvent fortement pénaliser ou influencer le modèle.

Passing et Bablok ont proposé une méthode qui permet de s'affranchir de ces hypothèses : les deux variables sont supposées comme comportant une part d'aléatoire (représentant l'erreur de mesure et la variabilité de la distribution de l'élément mesuré dans le milieu), sans qu'il soit fait d'hypothèse sur leur distribution, sinon qu'elle est identique. On pose alors :

- $y(i) = a + b \times x(i) + \eta(i)$,
- $x(i) = A + B \times y(i) + \xi(i)$,

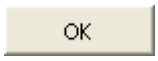
avec η et ξ qui suivent la même distribution. La méthode de Passing et Bablok permet de calculer les coefficients a et b (d'où A et B par les relations $B = \frac{1}{b}$ et $A = -\frac{1}{b}$) ainsi qu'un intervalle de confiance autour de ces valeurs. L'étude de ces valeurs permet de comparer les méthodes. Si elles sont très proches, on aura naturellement b proche de 1 et a proche de 0.

Par ailleurs, Passing et Bablok proposent un test de linéarité afin de vérifier que la relation entre les deux méthodes de mesure est stable sur le domaine d'étude. Ce test s'appuie sur une

statistique CUSUM, qui suit une distribution de Kolmogorov. XLSTAT fournit la statistique, la valeur critique pour le seuil de signification choisi, et la p-value associée à la statistique.

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas (mode colonnes), XLSTAT considère que les sites sont en lignes et les objets/variables en colonnes. Si la flèche est vers la droite (mode lignes), XLSTAT considère que les objets/variables sont en lignes et les sites en colonnes.

Onglet **Général** :

Méthode Y : sélectionnez les données qui correspondent aux $y(i)$ dans l'équation de régression. Si un libellé est présent en première position, veillez à ce que l'option « Libellé des variables » soit bien activée.

Méthode X : sélectionnez les données qui correspondent aux $x(i)$ dans l'équation de régression. Si un libellé est présent en première position, veillez à ce que l'option « Libellé des variables » soit bien activée.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des variables : activez cette option si, en mode colonnes, la première ligne des données sélectionnées contient un libellé, ou si en mode lignes, la première colonne des données sélectionnées contient un libellé.

Onglet **Options** :

Intervalle de confiance (%) : entrez la taille des intervalles de confiance (valeur par défaut : 95 %). La valeur correspondant à 100 moins le pourcentage entré est utilisée comme seuil de signification pour le test de linéarité.

Méthode :

- **Part I : même échelle** : cette méthode d'estimation est la première méthode développée par Passing et Bablok (1983). Elle doit être utilisée lorsque les deux méthodes sont sur une échelle identique et qu'elles évoluent dans le même sens (corrélation positive entre X et Y).
- **Part III : échelle différente** : cette méthode d'estimation développée par Bablok *et al.* en 1988 est une amélioration de la méthode appelée *Part I*. Elle est plus robuste et permet de comparer deux méthodes sur des échelles différentes avec éventuellement une corrélation négative entre X et Y.

Onglet **Données manquantes** :

Ne pas accepter les valeurs manquantes : activez cette option pour que XLSTAT empêche la poursuite des calculs si des valeurs manquantes sont détectées.

Supprimer les observations : activez cette option pour supprimer les observations comportant des données manquantes.

Onglet **Sorties** :

Statistiques descriptives : activez cette option pour afficher les statistiques descriptives pour les deux méthodes.

Onglet **Graphiques** :

Prédictions et résidus : activez cette option pour afficher les prédictions et les résidus pour l'ensemble des observations.

Résultats

Statistiques descriptives : dans un premier tableau vous trouverez les statistiques pour les deux méthodes étudiées.

Coefficients du modèle : dans ce tableau sont affichés les coefficients a et b du modèle, ainsi que l'intervalle de confiance autour de ces valeurs.

Prédictions et résidus : dans ce tableau sont affichés pour chaque observation, la valeur de X, la valeur de Y, la prédiction du modèle, le résidu et le résidu perpendiculaire (la distance à la droite de régression par projection orthogonale).

Les graphiques permettent de visualiser le modèle avec la droite de régression, les observations, et le modèle $Y = X$ (correspondant à la bissectrice du plan) et l'intervalle de confiance associé. Ce dernier étant calculé suivant la méthode de la régression linéaire classique sur la base d'un RMSE obtenu d'après le modèle de Passing et Bablok. Cela permet de visualiser l'éloignement éventuel du modèle par rapport à l'hypothèse que les méthodes sont identiques.

Exemple

Un exemple de comparaison de méthodes avec la régression de Passing et Bablok est disponible sur le Centre d'aide XLSTAT :

http://www.xlstat.com/demo-passingf.htm

Bibliographie

Passing H. and Bablok W. (1983). A new biometrical procedure for testing the equality of measurements from two different analytical methods. Application of linear regression procedures for method comparison studies in Clinical Chemistry, Part I. *Journal of Clinical Chemistry and Clinical Biochemistry*, **21**, 709-720.

Bablok, W., Passing, H., Bender, R., & Schneider, B. (1988). A general regression procedure for method transformation. Application of linear regression procedures for method comparison studies in clinical chemistry, Part III. *Clinical Chemistry and Laboratory Medicine*, **26(11)**, 783-790.

Régression de Deming

Utilisez cet outil pour comparer deux méthodes de mesure en supposant que X et Y contiennent toutes les deux une part d'erreur.

Dans cette section :

Description

Deming (1943) a développé une méthode de régression qui permet de comparer deux méthodes de mesure (par exemple, deux techniques pour mesurer la concentration d'un analyte), qui suppose que l'erreur de mesure soit présente aussi bien dans X et dans Y. Il s'affranchit des hypothèses lourdes et ici inappropriées de la régression linéaire simple classique. Pour rappel ces hypothèses sont :

- la variable explicative, X dans le modèle $y(i) = a + bx(i) + \epsilon(i)$, est déterministe (pas d'erreur de mesure),
- la variable Y suit une loi normale de moyenne aX
- la variance de l'erreur est constante.

Par ailleurs, les valeurs extrêmes peuvent fortement pénaliser ou influencer le modèle.

Deming a proposé une méthode qui permet de s'affranchir de ces hypothèses : les deux variables sont supposées comme comportant une part d'aléatoire (représentant l'erreur de mesure). On pose alors :

- $y(i) = y(i)^* + \epsilon(i)$
- $x(i) = x(i)^* + \eta(i)$

On suppose que les variables $(y(i), x(i))$ sont des mesures non exactes des vraies valeurs $(y(i)^*, x(i)^*)$ avec des termes d'erreur indépendants. On suppose néanmoins que le rapport des variances est connu :

$$\delta = \sigma^2(\eta)/\sigma^2(\epsilon)$$

XLSTAT-Life vous permet de fixer ces variances.

On recherche alors la droite permettant le « meilleur ajustement » $y^* = a + bx^*$, de manière à ce que la somme des carrés des résidus du modèle soit minimisée.

La méthode de Deming permet de calculer les coefficients a et b, ainsi qu'un intervalle de confiance autour de ces valeurs. L'étude de ces valeurs permet de comparer les méthodes. Si elles sont très proches, on aura naturellement b proche de 1 et a proche de 0.

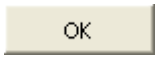
La regression de Deming peut prendre deux formes :

- La regression simple de Deming : Les termes d'erreur sont constants. L'estimation est très simple grâce à une formule directe.(Deming, 1943).
- La regression pondérée de Deming : Dans ce cas, on peut supposer que les erreurs sont proportionnelles. Cette approche est basée sur un algorithme iteratif (Linnet, 1990).

Les intervalles de confiance pour la constante et le coefficient de pente sont complexes à obtenir. XLSTAT-Life utilise le jackknife afin de les calculer comme indiqué dans Linnet (1993).

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas (mode colonnes), XLSTAT considère que les sites sont en lignes et les objets/variables en colonnes. Si la flèche est vers la droite (mode lignes), XLSTAT considère que les objets/variables sont en lignes et les sites en colonnes.

Onglet **Général**:

Y : sélectionnez les données qui correspondent aux $y(i)$ dans l'équation de régression. Si un libellé est présent en première position, veillez à ce que l'option « Libellé des variables » soit bien activée.

X : sélectionnez les données qui correspondent aux $x(i)$ dans l'équation de régression. Si un libellé est présent en première position, veillez à ce que l'option « Libellé des variables » soit bien activée.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des variables : activez cette option si, en mode colonnes, la première ligne des données sélectionnées contient un libellé, ou si en mode lignes, la première colonne des données sélectionnées contient un libellé.

Erreurs constantes : activez cette option si les erreurs de X et de Y sont supposées constantes.

Erreur proportionnelles : activez cette option si les erreurs de X et de Y sont supposées proportionnelles. On est alors dans le cas de la régression de Deming pondérée.

Onglet **Options**:

Intervalle de confiance (%) : entrez la taille des intervalles de confiance (valeur par défaut : 95 %). La valeur correspondant à 100 moins le pourcentage entré est utilisée comme seuil de signification pour le test de linéarité.

Variance de l'erreur sur X : entrez la variance de l'erreur de mesure sur X.

Variance de l'erreur sur Y : entrez la variance de l'erreur de mesure sur Y.

Onglet **Données manquantes** :

Ne pas accepter les valeurs manquantes : activez cette option pour que XLSTAT empêche la poursuite des calculs si des valeurs manquantes sont détectées.

Supprimer les observations : activez cette option pour supprimer les observations comportant des données manquantes.

Onglet **Sorties**:

Statistiques descriptives : activez cette option pour afficher les statistiques descriptives pour les deux méthodes.

Onglet **Graphiques** :

Prédictions et résidus : activez cette option pour afficher les prédictions et les résidus pour l'ensemble des observations.

Résultats

Statistiques descriptives : dans un premier tableau vous trouverez les statistiques pour les deux méthodes étudiées.

Coefficients du modèle : dans ce tableau sont affichés les coefficients a et b du modèle, ainsi que l'intervalle confiance autour de ces valeurs.

Prédictions et résidus : dans ce tableau sont affichés pour chaque observation, la valeur de X, la valeur de Y, la prédiction du modèle et le résidu.

Les graphiques permettent de visualiser le modèle avec la droite de régression, les observations, et le modèle $Y = X$ (correspondant à la bissectrice du plan) et l'intervalle de confiance associé, ce dernier étant calculé suivant la méthode de la régression linéaire classique sur la base d'un RMSE obtenu d'après le modèle de Deming. Cela permet de visualiser l'éloignement éventuel du modèle par rapport à l'hypothèse que les méthodes sont identiques.

Exemple

Un exemple de comparaison de méthodes avec la régression de Deming est disponible sur le Centre d'aide XLSTAT :

http://www.xlstat.com/demo-demingf.htm

Bibliographie

Deming, W. E. (1943). *Statistical adjustment of data*. Wiley, NY (Dover Publications edition, 1985).

Linnet K. (1990). Estimation of the Linear Relationship between the Measurements of Two Methods with Proportional Errors. *Statistics in Medicine,* Vol. 9, 1463-1473.

Linnet K. (1993). Evaluation of Regression Procedures for Method Comparison Studies. *Clin.Chem.* Vol. **39(3)**, 424-432.

Graphiques de Youden

Utilisez cet outil pour créer un graphique de Youden après avoir recueilli des résultats pour comparer deux matériels A et B évalués chacun par une série de laboratoires.

Dans cette section :

Description

Youden (1959) a développé une technique de représentation de données effectuées par N laboratoires pour deux matériels A et B similaires (ils peuvent être identiques lorsque l'on veut comparer des méthodes de mesure, ou différents mais attendus comme donnant des mesures identiques). Dans son article de 1959, Youden insiste sur la nécessité que la méthode soit simple, afin que ne soit pas nécessaire l'intervention d'un expert en statistique. L'objectif est ici d'identifier simplement quels laboratoires posent problème, soit parce que les deux mesures effectuées présentent un écart anormal, soit parce que les deux mesures sont trop différentes de ce qui est obtenu par d'autres laboratoires. Tant la variabilité inter-laboratoires que la variabilité intra-laboratoire sont ici analysées.

Le résultat est un graphique présentant les mesures pour le matériel A sur l'axe des abscisses et pour le matériel B sur l'axe des ordonnées. Un cercle, dans la version décrite par Youden, est ensuite affiché afin de pouvoir identifier des valeurs suspectes, celles qui sont à l'extérieur du cercle.

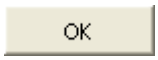
A l'origine, Youden pose comme pré-requis que les 2 matériels testés soient assez proches (deux techniques de mesure équivalentes, deux échantillons prélevés à deux endroits proches) et que les mesures soient sur des échelles identiques. Néanmoins, si ce n'est pas le cas, l'utilisateur de XLSTAT peut demander à ce que les données soient centrées réduites. Pour centrer-réduire les deux échantillons, l'utilisateur a la possibilité de choisir entre une standardisation classique (basée sur la moyenne arithmétique et la variance sans biais), ou une standardisation basée sur des statistiques robustes, telles que décrites dans la norme ISO 13528-2015-10 (algorithme A). Alternativement, l'utilisateur de XLSTAT pourra choisir l'une des deux méthodes d'affichage qui permettent de s'affranchir de la standardisation des données. XLSTAT propose donc trois types de représentation :

- Cercle : la technique proposée par Youden est directement appliquée.
- Ellipse : XLSTAT affiche une ellipse autour du nuage de points. Si le choix des statistiques robustes a été fait, la covariance robuste est utilisée pour le calcul de l'ellipse (Maronna, 2019). Le calcul de l'ellipse peut être fait en utilisant un intervalle de confiance s'appuyant sur la distribution de Fisher ou sur celle du Khi^2.

- Boîtes : XLSTAT affichera des rectangles qui sur chaque axe entourent la moyenne d'un intervalle de 2x2 et/ou 2x3 fois l'écart-type mesuré.

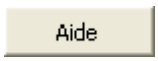
Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.



: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas (mode colonnes), XLSTAT considère que les sites sont en lignes et les objets/variables en colonnes. Si la flèche est vers la droite (mode lignes), XLSTAT considère que les objets/variables sont en lignes et les sites en colonnes.

Onglet **Général**:

Echantillon A : sélectionnez les données qui correspondent aux mesures effectuées par chacun des laboratoires pour l'échantillon A. Si un libellé est présent en première position, veillez à ce que l'option « Libellé des variables » soit bien activée.

Echantillon B : sélectionnez les données qui correspondent aux mesures effectuées par chacun des laboratoires pour l'échantillon B. Si un libellé est présent en première position, veillez à ce que l'option « Libellé des variables » soit bien activée.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des variables : activez cette option si, en mode colonnes, la première ligne des données sélectionnées contient un libellé, ou si en mode lignes, la première colonne des données sélectionnées contient un libellé.

Libellés des observations : activez cette option si vous voulez utiliser des libellés d'observations pour l'affichage des résultats. Si l'option « Libellés des variables » est activée, la première cellule de la sélection doit comprendre un en-tête. Si vous n'activez pas cette option, des libellés seront automatiquement créés (Obs1, Obs2, …).

Poids : activez cette option si vous voulez identifier des laboratoires dont les résultats seront affichés mais qui n'entreront pas dans le calcul des statistiques telles que la moyenne et l'écart-type. Il vous suffit de mettre la valeur 0 pour que le laboratoire correspondant soit éliminé des calculs.

Onglet **Options**:

Type de graphique :

- **Cercle** : activez cette option pour afficher un cercle.
- **Ellipse** : activez cette option pour afficher une ellipse. Vous avez alors le choix entre deux approches, celle qui consiste à utiliser la distribution de Fisher ou celle du Khi^2.
- **Boîtes** : activez cette option pour afficher une ou deux "boîtes" de confiance l'une située à 2 écart-types des moyennes de chaque mesure, l'autre située à 3 écart-types des moyennes de chaque mesure.

Intervalle de confiance (%) : entrez la taille des intervalles de confiance (valeur par défaut : 95 %). La valeur correspondant à 100 moins le pourcentage entré est utilisée comme seuil de signification pour le test de linéarité.

Standardiser : activez cette option pour centrer-réduire les données.

Statistiques robustes : activez cette option pour utiliser des statistiques robustes pour les moyennes, l'écart-type et pour le cas des ellipses la covariance.

Onglet **Données manquantes** :

Ne pas accepter les valeurs manquantes : activez cette option pour que XLSTAT empêche la poursuite des calculs si des valeurs manquantes sont détectées.

Supprimer les observations : activez cette option pour supprimer les observations comportant des données manquantes.

Onglet **Sorties**:

Statistiques descriptives : activez cette option pour afficher les statistiques descriptives pour les deux méthodes.

Résultats

Statistiques descriptives : dans un premier tableau vous trouverez les statistiques pour les deux méthodes étudiées.

Données et différences : dans ce tableau sont affichées les données (centrées réduites si l'option correspondante a été activée) des échantillons A et B, ainsi que les différences et les différences centrées absolues (on retire la différence des moyennes à la différence observée). La moyenne des |D| est utilisée pour le calcul du cercle de Youden.

Statistiques robustes : si les statistiques robustes ont été demandées, elles sont affichées.

Graphique de Youden : sur ce graphique sont affichées les données (centrées-réduites si l'option correspondante a été activée). Si les données ne sont pas centrées-réduites, une droite verticale et une droite horizontale passant par le point moyen sont affichées. La droite passant par la moyenne des deux échantillons (l'origine si les données sont centrées-réduites) et présentant un angle de 45° (option cercle) ou passant par l'axe de l'ellipse de pente positive (option ellipse) est affichée. Dans le cas où les boîtes ont été demandées ces dernières sont affichées sur le graphique.

Exemple

Un exemple de comparaison de mesures faites par 29 laboratoires pour deux échantillons faite en utilisant le graphique de Youden est disponible sur le Centre d'aide XLSTAT :

http://www.xlstat.com/demo-youdenf.htm

Bibliographie

Gnanadesikan R. and Kettenring J. R. (1972). Robust estimates, residuals, and outlier detection with multiresponse data. *Biometrics*, **28(1)**, 81-124.

ISO (2015). ISO 13528-2015-10, Statistical methods for use in proficiency testing by interlaboratory comparison, Second edition.

Maronna, R. A., Douglas Martin R., Yohai V.J. and Salibián-Barrera M. (2019). Robust Statistics Theory and Methods (with R), 2nd edition. Wiley, NJ.

Youden W.J. (1959). Graphical Diagnosis of Interlaboratory Test Results, *Journal of Quality Technology*, **15(11)**, 133-137.

Analyse d'effets de dose

Utilisez cette fonction pour modéliser les effets d'une dose sur une variable réponse, en prenant éventuellement en compte un effet de mortalité naturelle.

Dans cette section :

Description

Cet outil s'appuie sur la régression logistique (modèles Logit, Probit, Log- log complémentaire, Gompertz) pour modéliser l'impact de doses de composants chimiques (par exemple un médicament, un produit phytosanitaire) sur un phénomène binaire (guérison ou non, mort ou non).

Plus d'information sur la régression logistique est disponible dans la section de l'aide dédiée à ce sujet.

Mortalité naturelle

Cet outil permet de prendre en compte la mortalité naturelle afin de modéliser plus précisément le phénomène étudié. En effet, si l'on considère une expérience réalisée sur des insectes, certains périront en raison de la dose injectée, d'autres en raison d'un autre phénomène. L'ensemble de ces phénomènes connexes n'est pas intéressant pour l'expérience concernant les effets de dose, mais il peut être pris en compte. Si p est la probabilité issue d'un modèle de régression logistique correspondant uniquement à l'effet de la dose, et si m est la mortalité naturelle, alors la probabilité observée pour que l'insecte succombe est :

$$P(obs) = m + (1 - m) \times p$$

La formule d'Abbott (Finney, 1971) s'écrit

$$p = \frac{(P(obs) - m)}{(1 - m)}$$

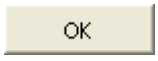
La mortalité naturelle m peut être entrée par l'utilisateur parce que connu grâce à des expériences préalables, ou déterminée par XLSTAT.

ED 50, ED 90, ED 99

XLSTAT permet de calculé les doses ED50 (ou dose médiane), ED90 et ED99 qui correspondent aux doses entraînant un effet sur respectivement 50%, 90% et 99% de la population.

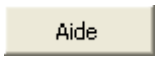
Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.



: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.

Onglet **Général** :

Variables dépendantes :

Variable(s) réponse : sélectionnez la ou les variables réponse que vous souhaitez modéliser. Si plusieurs variables sont sélectionnées, XLSTAT fera les calculs pour chacune des variables indépendamment. Si des en-têtes de colonnes ont été sélectionnés, veuillez vérifier que l'option « Libellés des variables » est activée.

Type de réponse : choisissez le type de variable réponse que vous avez sélectionné :

- Variable binaire : si vous sélectionnez cette option, vous devez sélectionner une variable contenant exactement deux valeurs distinctes. Si la variable est constituée de 0 et de 1, XLSTAT fera en sorte que les probabilités élevées du modèle correspondent à la catégorie 1, et que les probabilités faibles correspondent à la catégorie 0. Si la variable comprend deux autres valeurs (par exemple Oui / Non), à la première catégorie rencontrée correspondront les faibles probabilités et à la seconde les probabilités élevées.

- Somme de variables binaires : si votre variable réponse correspond à une somme de variables binaires, elle doit être de type numérique et contenir le nombre d'événements positifs (événement 1) parmi tous ceux observés. La variable correspondant au nombre total d'événements observés pour cette observation (événements 1 et 0 combinés) doit alors être sélectionnée dans le champ « poids des observations ». Ce cas correspond par exemple à une expérience où l'on administre une dose D d'un médicament (D est la variable explicative) à 50 patients (50 est la valeur du poids des observations), et où l'on observe que 40 sont guéris sous l'effet de la dose (40 correspond à la valeur de la variable réponse).

Variables explicatives :

Quantitatives : activez cette option si vous voulez inclure une ou plusieurs variables explicatives quantitatives dans le modèle. Sélectionnez alors la ou les variables correspondantes sur la feuille Excel. Les données sélectionnées doivent être de type numérique. Si le libellé des variables a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

Qualitatives : activez cette option si vous voulez inclure une ou plusieurs variables explicatives qualitatives dans le modèle. Sélectionnez alors la ou les variables correspondantes sur la feuille Excel. Les données sélectionnées peuvent être de tout type, mais les données numériques sont automatiquement considérées comme nominales. Si le libellé des variables a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

Modèle : choisissez le type de fonction à utiliser (voir description).

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des variables : activez cette option si la première ligne des données sélectionnées (variables dépendantes et explicatives, libellés des observations, poids) contient un libellé.

Libellés des observations : activez cette option si vous voulez utiliser des libellés d'observations pour l'affichage des résultats. Si l'option « Libellés des variables » est activée, la première cellule de la sélection doit comprendre un en-tête. Si vous n'activez pas cette option, des libellés seront automatiquement créés (Obs1, Obs2, …).

Poids des observations : ce champ est à remplir impérativement si l'option « somme de binaires » a été choisie. Sinon ce champ n'est pas actif. Si un en- tête de colonne a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

Onglet **Options** :

Méthode de Firth : activez cette option pour utiliser la vraisemblance pénalisée de Firth (voir description).

Intervalle de confiance (%) : entrez l'étendue en pourcentage de l'intervalle de confiance à utiliser pour les différents tests, et pour le calcul des intervalles de confiance autour des paramètres et des prédictions. Valeur par défaut : 95%.

Conditions d'arrêt :

- **Itérations** : entrez le nombre maximal d'itérations pour l'algorithme de Newton-Raphson. Les calculs sont interrompus dès que le nombre maximal d'itérations est dépassé. Valeur par défaut : 100.
- **Convergence** : entrez la valeur seuil d'évolution maximale de log vraisemblance d'une itération à l'autre, qui une fois atteinte permet de considérer que l'algorithme a convergé. Valeur par défaut : 0,000001.

Utiliser le logarithme : activez cette option pour utiliser le logarithme des variables quantitatives dans le modèle.

Paramètre de mortalité naturelle : activez cette option pour inclure un paramètre de mortalité naturelle dans le modèle.

- **Optimisé** : choisissez cette option pour que XLSTAT trouve la valeur du paramètre maximisant la vraisemblance du modèle.
- **Défini par l'utilisateur** : entrez la valeur de la mortalité naturelle. Cette valeur doit être comprise entre 0 et 0.9. Valeur par défaut : 0,1.

Onglet **Validation** :

Validation : activez cette option si vous souhaitez utiliser une partie des données sélectionnées pour valider le modèle.

Jeu de validation : choisissez l'une des options pour définir le mode de sélection des observations utilisées pour la validation :

- **Aléatoire** : les observations sont sélectionnées de manière aléatoire. Le « Nombre d'observations » doit alors être saisi.
- **N dernières lignes** : les N dernières observations sont sélectionnées pour la validation. Le « Nombre d'observations » N doit alors être saisi.
- **N premières lignes** : les N premières observations sont sélectionnées pour la validation. Le « Nombre d'observations » N doit alors être saisi.

- **Variable de groupe** : si vous choisissez cette option, vous devez ensuite sélectionner une variable indicatrice composée de 0 pour les observations à utiliser pour le calcul du modèle, et de 1 pour les observations à utiliser pour la validation du modèle.

Onglet **Prédiction** :

Prédiction : activez cette option si vous souhaitez sélectionner des données à utiliser en mode prédiction. Si vous activez cette option, vous devez veiller à ce que les données de prédiction soient organisées comme les données d'estimation : mêmes variables, même ordre dans les sélections. En revanche vous ne devez pas sélectionner de libellés de variables : la première ligne des sélections décrites ci-dessous doit être une ligne de données.

Quantitatives : activez cette option pour sélectionner la ou les variables quantitatives explicatives. La première ligne ne doit pas comprendre d'en- tête.

Qualitatives : activez cette option pour sélectionner la ou les variables qualitatives explicatives. La première ligne ne doit pas comprendre d'en-tête.

Libellés des observations : activez cette option si vous voulez utiliser des libellés d'observations disponibles sur une feuille Excel pour l'affichage des résultats. La première ligne ne doit pas comprendre d'en-tête. Si vous n'activez pas cette option, des libellés seront automatiquement créés (PredObs1, PredObs2, …).

Libellés des variables : activez cette option si la première ligne des données sélectionnées (variables explicatives, libellés des observations) contient un libellé.

Onglet **Données manquantes** :

Supprimer les observations : activez cette option pour supprimer les observations comportant des données manquantes.

Estimer les données manquantes : activez cette option pour estimer les données manquantes avant le début des calculs.

- **Moyenne ou mode** : activez cette option pour estimer les données manquantes en utilisant la moyenne (variables quantitatives) ou le mode (variables qualitatives) pour les variables correspondantes.

- **Plus proche voisin** : activez cette option pour estimer les données manquantes d'une observation en recherchant le plus proche voisin de l'observation.

Onglet **Sorties** :

Statistiques descriptives : activez cette option pour afficher les statistiques descriptives pour les variables sélectionnées.

Corrélations : activez cette option pour afficher la matrice de corrélations des variables explicatives.

Coefficients d'ajustement : activez cette option pour afficher le tableau des statistiques d'ajustement du modèle.

Analyse de type III : activez cette option pour afficher le tableau d'analyse de la variable de type III.

Coefficients du modèle : activez cette option pour afficher le tableau des coefficients du modèle. Optionnellement les **intervalles de confiance** de type « *profile likelihood* » peuvent être calculés (voir description).

Coefficients normalisés : activez cette option pour afficher les paramètres normalisés du modèle (coefficients bêta).

Equation : activez cette option pour afficher explicitement l'équation du modèle.

Prédictions et résidus : activez cette option pour afficher les prédictions et les résidus pour l'ensemble des observations.

Analyse des probabilités : si une seule variable explicative a été sélectionnée, activez cette option pour que XLSTAT calcule la valeur de la variable explicative correspondant à divers niveaux de probabilité.

Onglet **Graphiques** :

Graphiques de régression : activez cette option pour afficher les graphiques de régression :

- **Coefficients normalisés** : activez cette option pour afficher sur un graphique les paramètres normalisés du modèle avec leur intervalle de confiance.
- **Prédictions** : activez cette option pour afficher la courbe de régression.
- **Intervalles de confiance** : activez cette option pour afficher les intervalles de confiance sur les graphiques (1) et (4).

Résultats

XLSTAT propose un nombre important de tableaux et de graphiques afin de faciliter l'analyse et l'interprétation des résultats.

Statistiques descriptives : le tableau de statistiques descriptives présente pour toutes les variables sélectionnées des statistiques simples. Pour les variables quantitatives Sont affichés le nombre d'observations, le nombre de données manquantes, le nombre de données non manquantes, la moyenne, et l'écart-type (non biaisé). Pour les variables qualitatives, dont la variable dépendante, sont affichées les modalités leurs effectifs et pourcentage respectifs.

Matrice de corrélation : dans ce tableau sont affichées les corrélations entre les variables explicatives.

Correspondance entre les modalités de la variable réponse et les probabilités : ce tableau permet de visualiser à quelles modalités de la variable dépendante ont été affectées les probabilités 0 et 1.

Coefficients d'ajustement : dans ce tableau est affichée une série de statistiques pour le modèle indépendant (correspondant au cas où la combinaison linéaire des variables explicatives se réduit à une constante) et pour le modèle ajusté.

- **Observations** : le nombre total d'observations prises en compte (somme des poids des observations) ;
- **Somme des poids** : le nombre total d'observations prises en compte (somme des poids des observations multipliés par les poids dans la régression) ;
- **DDL** : degrés de liberté ;
- **-2 Log(Vrais.)** : le logarithme de la fonction de vraisemblance associée au modèle;
- **R² (McFadden) :** coefficient compris comme le R^2 entre 0 et 1 qui mesure le bon ajustement du modèle. Ce coefficient est égal à 1 moins le rapport de la vraisemblance du modèle ajusté sur la vraisemblance du modèle indépendant.
- **R²(Cox et Snell) :** coefficient compris comme le R^2 entre 0 et 1 qui mesure le bon ajustement du modèle. Ce coefficient est égal à 1 moins le rapport de la vraisemblance du modèle ajusté sur la vraisemblance du modèle indépendant, le rapport étant porté à l'exposant $\frac{2}{S_w}$, où S_w est la somme des poids ;
- **R²(Nagelkerke) :** coefficient compris comme le R^2 entre 0 et 1 qui mesure le bon ajustement du modèle. Ce coefficient est égal au rapport du R^2 de Cox et Snell, divisé par 1 moins le la vraisemblance du modèle indépendant portée à l'exposant $\frac{2}{S_w}$;
- **AIC** : le critère d'information d'Akaike (Akaike's Information Criterion) ;
- **SBC** : le critère bayésien de Schwarz (Schwarz's Bayesian Criterion).

Test de l'hypothèse nulle H0 : **Y=p0** : l'hypothèse H_0 correspond au modèle indépendant qui donne la probabilité p_0 quelques soient les valeurs des variables explicatives ; on cherche à vérifier si le modèle ajusté est significativement plus performant que ce modèle. Trois tests sont proposés : le test du rapport des vraisemblance (-2 Log(Vrais.)), le test du Score, et le test test de Wald. Les trois statistiques suivent une loi du χ^2 dont les degrés de liberté sont indiqués.

Analyse de Type III : ce tableau n'a d'intérêt que s'il y a plus d'une variable explicative. On test ici le modèle ajusté contre un test dont on aurait retiré la variable de la ligne du tableau en question. Si la probabilité $Pr > LR$ est inférieur à un seul de signification que l'on se fixe (typiquement 0.05), alors la contribution de la variable à l'ajustement du modèle est significative. Sinon, elle peut être retirée du modèle.

Paramètres du modèle : pour la constante du modèle et pour chaque variable sont affichés l'estimation du paramètre, l'écart-type correspondant, le χ^2 de Wald, la p-value correspondante, ainsi que l'intervalle de confiance. Si l'option correspondante a été activée, les intervalles « *profile likelihood* » sont aussi affichés.

L'**équation du modèle** est ensuite affichée pour faciliter la lecture ou la réutilisation du modèle.

Le tableau des **coefficients normalisés** (aussi appelés coefficients bêta) permet de comparer le poids relatif des variables. Plus la valeur absolue d'un coefficient est élevée, plus le poids de la variable correspondante est important. Lorsque l'intervalle de confiance autour des coefficients normalisés comprend la valeur 0 (cela est facilement visible sur le graphique des coefficients normalisés), le poids d'une variable dans le modèle n'est pas significatif.

Dans le tableau des **prédictions et résidus** sont donnés pour chaque observation, son poids, la valeur de la variable explicative qualitative s'il n'y en a qu'une, la valeur observée de la variable dépendante, la prédiction du modèle, les mêmes valeurs divisées par le poids, les résidus standardisés, ainsi qu'un intervalle de confiance.

Le tableau d'**analyse des probabilités** n'est affiché que si une seule variable explicative quantitative a été sélectionnée. Il permet de visualiser à quel niveau de la variable explicative correspond une probabilité donnée.

Exemple

Un exemple d'analyse d'effets de dose est disponible sur le Centre d'aide XLSTAT à l'adresse

http://www.xlstat.com/demo-dosef.htm

Bibliographie

Abbott W.S. (1925). A method for computing the effectiveness of an insecticide. Jour. Econ. Entomol. 18 : 265-267.

Agresti A. (1990). Categorical Data Analysis. John Wiley and Sons, New York.

Finney D.J. (1971). Probit Analysis, 3rd Edition. Cambridge, London and New-York.

Firth D (1993). Bias reduction of maximum likelihood estimates. *Biometrika*, **80**, 27-38.

Furnival G. M. and Wilson R.W. Jr. (1974). Regressions by leaps and bounds. *Technometrics*, **16** (4), 499-511.

Heinze G. and Schemper M. (2002). A solution to the problem of separation in logistic regression.*Statistics in Medicine*, **21**, 2409-2419.

Hosmer D.W. and Lemeshow S. (2000). Applied Logistic Regression, Second Edition. John Wiley and Sons, New York.

Lawless J.F. and Singhal K. (1978). Efficient screening of nonnormal regression Models. *Biometrics*, **34**, 318-327.

Tallarida R.J. (2000). Drug Synergism & Dose-Effect Data Analysis. CRC/Chapman and Hall, Boca Raton.

Venzon, D. J. and Moolgavkar S. H. (1988). A method for computing profile likelihood based confidence intervals. *Applied Statistics*, **37**, 87-94.

Régression logistique à 4 ou 5 paramètres et courbes parallèles

Utilisez cet outil pour modéliser l'effet d'une variable quantitative sur une variable réponse (densité optique, concentration, ect.), en utilisant le modèle logistique à 4 ou 5 paramètres, et en tenant éventuellement compte de contraintes liées à l'existence d'un échantillon standard.

Dans cette section :

Description

Le modèle logistique à 4 paramètres est donné par l'équation suivante :

$$y = a + \frac{d-a}{1+(\frac{x}{c})^b} \tag{1}$$

où a, b, c, d sont les paramètres du modèle, et x est la variable explicative, et y la variable réponse. Les paramètres a et d sont des paramètres d'asymptotes (a étant le minimum et d le maximum), et b est le paramètre de pente. Le paramètre c correspond à l'abscisse du point de mi-pente dont l'ordonnée est $(a+d)/2$. Lorsque a est inférieur à d, la courbe descend de d à a et, lorsque a est supérieur à d, la courbe monte de a à d.

Le modèle logistique à 5 paramètres est donné par l'équation suivante :

$$y = a + \frac{d-a}{\left[1+(\frac{x}{c})^b\right]^e} \tag{2}$$

Le paramètre e est un paramètre d'asymétrie.

Pour l'ajustement parallèle à 4 paramètres, le modèle utilisé est le suivant :

$$y = a + \frac{d-a}{1+(s0\cdot\frac{x}{c0}+s1\cdot\frac{x}{c1})^b} \tag{3}$$

où s0 vaut 1 si la donnée x provient de l'**échantillon standard**, et 0 sinon, et où s_1 vaut 1 si la donnée x provient de l'**échantillon étudié**, et 0 sinon. Ce modèle est dit sous contrainte, car pour l'estimation des paramètres a, b, et d, les valeurs obtenues pour l'échantillon standard sont prises en compte. De la description des paramètres ci-dessus, on comprend que ce

modèle génère deux courbes parallèles, dont la seule différence est la position, le décalage étant donné par $(c_1 - c_0)$. Si c_1 est supérieur à c_0, la courbe correspondant à l'échantillon étudié sera décalée à droite de la courbe correspondant à l'échantillon standard, et vice-versa.

Pour l'ajustement parallèle à 5 paramètres, le modèle utilisé est le suivant :

$$y = a + \frac{d - a}{(1 + (s0 \cdot \frac{x}{c0} + s1 \cdot \frac{x}{c1})^b)^e} \quad (4)$$

XLSTAT permet d'ajuster :

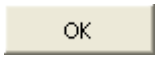
- Le modèle (1) ou (2), à un échantillon standard ou à un échantillon étudié,
- Le modèle (3) ou (4), à la fois à l'échantillon standard et à l'échantillon étudié.

Si l'utilisateur le souhaite, XLSTAT peut tester pour chaque échantillon (standard et étudié) si des valeurs extrêmes perturbent l'ajustement. Pour les modèles (1) ou (2), le test de Dixon est appliqué une fois le modèle ajusté. Si une valeur anormale est détectée, elle est supprimée, et le modèle est recalculé, et ainsi de suite jusqu'à ce que plus aucune valeur extrême ne soit détectée. Pour les modèles (3) et (4), on effectue d'abord un test de Dixon avec les modèles (1) ou (2) sur l'échantillon standard, puis sur l'échantillon étudié, puis les modèles (3) ou (4) sont ajustés au regroupement des deux échantillons sans les observations supprimées.

Pour les modèles (3) ou (4) un test de Fisher est effectué afin de déterminer si les paramètres a, b et d (et éventuellement es) obtenus avec les modèles (1) ou (2) ne sont pas significativement différents pour les deux échantillons pris séparément.

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les

variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.

Onglet **Général**:

Y / Variables dépendantes :

Quantitatives : sélectionnez la ou les variables réponse que vous souhaitez modéliser. Si plusieurs variables sont sélectionnées, XLSTAT fera les calculs pour chacune des variables indépendamment. Si des en-têtes de colonnes ont été sélectionnés, veuillez vérifier que l'option « Libellés des variables » est activée.

X / Variables explicatives : sélectionnez alors la ou les variables quantitatives explicatives sur la feuille Excel. Les données sélectionnées doivent être de type numérique. Si le libellé des variables a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

Modèle :

- **4PL** : activez cette option pour ajuster un modèle à 4 paramètres.
- **5PL** : activez cette option pour ajuster un modèle à 5 paramètres.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des variables : activez cette option si la première ligne des données sélectionnées (variables dépendantes et explicatives, libellés des observations, poids) contient un libellé.

Libellés des observations : activez cette option si vous voulez utiliser des libellés d'observations pour l'affichage des résultats. Si l'option « Libellés des variables » est activée, la première cellule de la sélection doit comprendre un en-tête. Si vous n'activez pas cette option, des libellés seront automatiquement créés (Obs1, Obs2, …).

Sous-échantillons : activez cette option si vous souhaitez distinguer parmi les données sélectionnées, un échantillon standard (identifiant 0) d'autres échantillons (identifiants 1,2…). Sélectionnez alors la colonne des identifiants. Si le libellé des variables a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

Onglet **Options**:

Valeurs de départ : activez cette option pour donner un point de départ à XLSTAT. Sélectionnez alors les cellules correspondant aux valeurs initiales des paramètres. Le nombre de lignes sélectionnées doit correspondre au nombre de paramètres.

Bornes des paramètres : activez cette option pour indiquer à XLSTAT une région possible pour l'ensemble des paramètres du modèle choisi. Vous devez alors sélectionner une plage de deux colonnes, celle de gauche correspondant aux bornes inférieures, et celle de droite aux bornes supérieures. Le nombre de lignes sélectionnées doit correspondre au nombre de paramètres.

Libellés des paramètres : activez cette option si vous voulez préciser les noms des paramètres. Au lieu d'afficher les noms génériques pr1, pr2, etc., pour les paramètres, XLSTAT affichera les résultats en utilisant les libellés sélectionnés. Le nombre de lignes sélectionnées doit correspondre au nombre de paramètres.

Conditions d'arrêt :

- **Itérations** : entrez le nombre maximal d'itérations pour l'algorithme d'ajustement. Les calculs sont interrompus dès que le nombre maximal d'itérations est dépassé. Valeur par défaut : 50.
- **Convergence** : entrez la valeur seuil d'évolution maximale de la somme des carrés des erreurs (SCE) d'une itération à l'autre, qui une fois atteinte permet de considérer que l'algorithme a convergé. Valeur par défaut : 0,0001.

Onglet **Données manquantes** :

Supprimer les observations : activez cette option pour supprimer les observations comportant des données manquantes.

Estimer les données manquantes : activez cette option pour estimer les données manquantes avant le début des calculs.

- **Moyenne ou mode** : activez cette option pour estimer les données manquantes en utilisant la moyenne (variables quantitatives) ou le mode (variables qualitatives) pour les variables correspondantes.
- **Plus proche voisin** : activez cette option pour estimer les données manquantes d'une observation en recherchant le plus proche voisin de l'observation.

Test de Dixon : activez cette option pour utiliser le test de Dixon pour supprimer les valeurs extrêmes de l'échantillon d'estimation.

Intervalles de confiance : activez cette option pour entre la taille de l'intervalle de confiance pour le test de Dixon.

Onglet **Sorties**:

Statistiques descriptives : activez cette option pour afficher les statistiques descriptives pour les variables sélectionnées.

Coefficients d'ajustement : activez cette option pour afficher le tableau des statistiques d'ajustement du modèle.

Paramètres du modèle : activez cette option pour afficher les valeurs des paramètres du modèle après ajustement.

Equation du modèle : activez cette option pour afficher l'équation du modèle.

Prédictions et résidus : activez cette option pour afficher les prédictions et les résidus pour l'ensemble des observations.

Onglet **Graphiques**:

Données et prédictions : activez cette option pour afficher le graphique des données observées et la courbe de la fonction ajustée.

- **Echelle logarithmique** : activez cette option pour que le graphique soit affiché sur l'échelle logarithmique.

Résidus : activez cette option pour afficher le diagramme en bâtons des résidus.

Résidus/Prédictions : activez cette option pour afficher le graphique des résidus en fonction des prédictions.

Résultats

Statistiques simples : le tableau de statistiques descriptives présente pour toutes les variables sélectionnées des statistiques simples : le nombre d'observations, le nombre de données manquantes, le nombre de données non manquantes, la moyenne, et l'écart-type (non biaisé).

Si aucun groupe ou un seul échantillon a été sélectionné, les résultats sont affichés pour le modèle choisi et pour cet échantillon. Si plusieurs sous- échantillons ont été définis (option sous-échantillons dans la boîte de dialogue), le modèle est d'abord ajusté à l'échantillon standard, puis chaque sous-échantillon est comparé à l'échantillon standard.

Test de Fisher du parallélisme entre les courbes : le test de Fisher est utilisé pour déterminer si on peut considérer que l'échantillon standard et l'échantillon étudié ont des paramètres a, b, d et e (e n'est utilisé que si le modèle à 5 paramètre a été retenu) significativement identiques ou non. Si la probabilité associée à la valeur F obtenue est inférieure au seuil de signification que l'on se fixe (5% par exemple), alors on peut considérer que les deux échantillons ont des paramètres a, b, d et e significativement différents.

Coefficients d'ajustement : dans ce tableau sont affichées les statistiques suivantes :

- le nombre d'observations ;
- le nombre de degrés de liberté (DDL) ;
- le coefficient de détermination R^2 ;
- la somme des carrés des erreurs (ou résidus) du modèle (SCE) ;
- la moyenne des carrés des erreurs (ou résidus) du modèle (MCE) ;
- la racine de la moyenne des carrés des erreurs (ou résidus) du modèle (RMCE) ;

Paramètres du modèle : dans ce tableau sont affichés les estimateurs de chacun des paramètres du modèle, ainsi que l'écart-type correspondant.

Prédictions et résidus : ce tableau donne pour chaque observation, la valeur de la variable d'échantillon, les données de départ, la valeur prédite pour le modèle, les résidus. Si des observations ont été supprimées suite au test de Dixon, elles sont affichées en gras.

Graphiques : sur le premier graphique sont figurées en bleu les données et la courbe correspondant à l'échantillon standard, et en rouge les données et la courbe correspondant à l'échantillon étudié. Sur le deuxième graphique sont affichées les valeurs observées pour la variable dépendante en fonction des valeurs prédites. Le troisième graphique correspond au diagramme en bâtons des résidus. Le dernier graphique correspond au graphique des résidus en fonction des prédictions.

Exemple

Un exemple de régression logistique à 4 paramètres est disponible sur le Centre d'aide XLSTAT :

http://www.xlstat.com/demo-4plf.htm

Bibliographie

Dixon W.J. (1953). Processing data for outliers, *Biometrics*, **9**, 74-89.

Tallarida R.J. (2000). Drug Synergism & Dose-Effect Data Analysis. CRC/Chapman & Hall, Boca Raton.

Expression différentielle

Utilisez cet outil pour détecter les caractères les plus différentiellement exprimés selon des variables explicatives au sein d'un tableau de données caractères/individus pouvant atteindre de très grandes dimensions.

Dans cette section :

Description

L'expression différentielle permet d'identifier les caractères (gènes, protéines, métabolites) les plus significativement affectés par des variables explicatives contrôlées (exemple : comparaison d'individus sains à des individus malades). Les données utilisées ont souvent une taille spectaculaire (= obtenues à haut débit). On parle aussi de données OMICS, en référence à des données recueillies à l'échelle du génome (genomics), du transcriptome (transcriptomics), du protéome (proteomics), du métabolome (metabolomics), etc.

La détection de caractères différentiellement exprimés met souvent en jeu des tests statistiques classiques. Cependant, le volume des données peut poser problème en termes de temps de calcul, de fiabilité statistique des résultats ainsi que de leur lisibilité. Des adaptations de ces outils sont par conséquent mises en œuvre afin de pallier ces problèmes.

Tests statistiques

Les tests statistiques proposés au sein de l'outil expression différentielle de XLSTAT sont des tests classiques, paramétriques ou non-paramétriques, qui sont documentés dans d'autres sections de l'aide : test t de Student, ANOVA, Mann-Whitney, Kruskal-Wallis. Un troisième test, moins classique, est disponible qui est basé sur l'approche bayésienne empirique dont une description est donnée ci-dessous.

Test Bayésien empirique

Cette méthode repose sur l'ajustement d'un modèle linéaire à chaque caractère j. Pour cela la méthode des moindres carrés est utilisée pour estimer l'ensemble des coeffcients $\hat{\alpha}_j$ tel que $$ E(y_j) \approx X \tilde\alpha_j, \quad j = 1, n, $$

où X est la matrice design contenant les valeurs d'un covarié pour les différentes observations sur la variable dépendante. La variance du k-ème coefficient est supposée vérifier la relation $$

var((\hat\alpha_j)*k) = \sigma^2_j((X^T X)^{-1}) $$

En remplaçant σ_j^2 par un estimé il est ensuite possible d'effectuer différents tests statistiques sur ces coefficients. Dans le cas précis du test bayésien empirique, la statistique-t modérée est définie comme : $$ t = \frac{\hat\beta_j}{\tilde s_j \sqrt{v_j}} $$

où $\hat{\beta}_j = C\hat{\alpha}_j$ est une combinaison linéaire des coefficients estimés, $v_j = C^T (X^T X)^{-1} C$ et $\tilde{s}$ est l'estimé posterior de la variance. On va supposer que la véritable variance du gène j, s'il n'est pas différentiellement exprimé, est tirée de la distribution : $$ \frac{1}{\sigma_j^2} \approx \frac{1}{d_0 s_0^2}\chi^2_{d_0} $$

où s_0 est la valeur attendue des variances réelles. De cette dernière équation, il est possible de démontrer que $$ \tilde s_j^2 = \frac{d_0 s_0^2 + d_j s_j^2}{d_0+d_j}. $$

Corrections post-hoc

La p-value représente le risque que l'on prend de se tromper en affirmant qu'un effet est statistiquement significatif. Effectuer un test en boucle un grand nombre de fois augmente le nombre de p-values calculées et par conséquent le risque de détecter des effets significatifs à tort. Avec un seuil de risque alpha de 5%, il est probable de détecter 5 p-values significatives par hasard sur 100 p-values calculées. En travaillant sur les données à haut-débit, on est souvent amené à tester par exemple l'effet d'une variable explicative sur l'expression de plusieurs milliers de gènes, impliquant ainsi le calcul de plusieurs milliers de p-values. Par conséquent, les p-values doivent être corrigées (= augmentées = pénalisées) à mesure que leur nombre augmente. XLSTAT propose 3 méthodes courantes de corrections :

Benjamini-Hochberg : cette procédure fait en sorte que les p-values augmentent en fonction de leur nombre et du taux de p-values non-significatives. Elle fait partie de la famille de correction type FDR (False Discovery Rate). Etant peu conservatrice (= peu sévère), elle est bien adaptée aux situations où l'on cherche à sélectionner un grand nombre de caractères potentiellement intéressants. Elle est très souvent utilisée dans les problématiques d'expression différentielle.

La p-value corrigée selon la procédure de Benjamini-Hochberg est définie de la sorte :

$$p_{BenjaminiHochberg} = min(p \times nbp/j, 1)$$

p étant la p-value d'origine, nbp le nombre de p-values calculées au total et j le rang de la p-value lorsque les p-values sont rangées par ordre croissant.

Benjamini-Yekutieli : cette procédure fait en sorte que les p-values augmentent en fonction de leur nombre et du taux de p-values non- significatives. Elle fait partie de la famille de correction type FDR (False Discovery Rate). En plus de la procédure de Benjamini-Hochberg, elle prend en compte une possible dépendance entre les éléments testés. Elle est par conséquent un peu plus conservatrice que la procédure précédente mais beaucoup moins que celle de Bonferroni.

La p-value corrigée selon la procédure de Benjamini-Yekutieli est définie de la sorte :

$$p_{BenjaminiYekutieli} = min[(p \times nbp \sum_{i=1}^{nbp} 1/i)/j, 1]$$

p étant la p-value d'origine, nbp le nombre de p-values calculées au total et j le rang de la p-value lorsque les p-values sont rangées par ordre croissant.

Bonferroni : les p-values n'augmentent qu'en fonction de leur nombre. Cette procédure est très conservatrice. Elle fait partie de la famille de correction type FWER (Familywise Error Rate). Elle est peu souvent utilisée dans les études d'expression différentielle. Elle s'avère utile lorsque l'utilisateur cherche à ne détecter qu'un nombre réduit de caractères différentiellement exprimés.

La p-value corrigée selon la procédure de Bonferroni est définie de la sorte :

$$p_{Bonferroni} = min(p \times nbp, 1)$$

p étant la p-value d'origine et nbp le nombre de p-values calculées.

Comparaisons multiples par paires

Suite à des ANOVA à un facteur et des tests de Kruskal-Wallis, il est possible de procéder à des tests de comparaisons multiples par caractère. XLSTAT propose différentes options :

Test de Tukey (HSD) : ce test est le plus utilisé (HSD : honestly significant difference).

Test de Fisher (LSD) : c'est un test de Student qui permet de tester l'hypothèse nulle que toute les moyennes pour les différentes modalités sont égales (LSD : least significant difference).

Test du t* de Bonferroni : dérivé du test de Student, il est un peu plus performant car il prend en compte le fait que plusieurs comparaisons sont effectuées simultanément. En conséquence, le niveau de signification du test est modifié suivant la formule suivante :

$$\alpha^{'} = \alpha/(g(g-1)/2)$$

où g est le nombre de modalités du facteur dont les modalités sont comparées.

Test de Dunn-Sidak : dérivé du test de Bonferroni, il est plus performant dans certaines situations.

$$\alpha' = 1 - (1-\alpha)^{2/[g(g-1)]}$$

Filtrage non spécifique

Avant de lancer les analyses, il est intéressant d'éliminer les caractères dont l'expression est peu variable à travers les individus. Le filtrage non- spécifique a deux avantages principaux :

- Il fait en sorte que le calcul se focalise moins sur les caractères probablement non exprimés différentiellement.
- Il limite les pénalisations post-hoc, puisque le nombre de p-values calculées est plus faible.

Deux méthodes sont disponibles dans XLSTAT :

- L'utilisateur indique un seuil de variabilité (écart interquartile ou écart type). Les caractères dont la variabilité est plus faible que ce seuil sont éliminés en amont des analyses.
- L'utilisateur spécifie un pourcentage de caractères avec une faible variabilité (écart interquartile ou écart type) à éliminer en amont des analyses.

Effets biologiques et effets statistiques : le volcano plot

Qui dit effet statistiquement significatif ne dit pas nécessairement effet biologique important. Un dispositif expérimental impliquant des mesures très précises, avec un très grand nombre de répétitions, peut être à l'origine de p-values faibles pour des différences biologiques pourtant infimes. Pour cette raison, il est toujours recommandé de « garder un œil » sur le biologique et de ne pas se fier exclusivement à ce que nous racontent les p-values. Le volcano plot est un nuage de points combinant effet statistique sur l'axe des ordonnées et effet biologique sur l'axe des abscisses pour une matrice de données caractères/individus. La seule contrainte est qu'il ne peut être appliqué que pour examiner les différences entre les modalités de variables qualitatives explicatives à deux modalités.

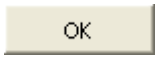
L'axe des ordonnées représente -log10(p-value). Cela facilite la lecture du graphique : les valeurs élevées représentent des effets significatifs et les valeurs faibles des effets non-significatifs.

XLSTAT propose deux manières de construire l'axe des abscisses, notamment :

- Différence entre la moyenne de la première modalité et la moyenne de la deuxième, pour chaque caractère. En général, on utilise ce format pour des données ayant subi une transformation d'échelle, type logarithmique ou racine.
- Le log en base 2 du ratio des moyennes des deux modalités : log2(moyenne1/moyenne2). Plutôt utilisé pour les données non-transformées.

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.

 : cliquez sur ce bouton pour effacer les sélections de données.

Onglet **Général** :

Tableau des Caractères/Individus : sélectionnez la matrice de données caractères/individus. Les données sélectionnées doivent être de type numérique.

Format des données :

Caractères en lignes : activez cette option si les caractères sont disposés en lignes et les individus (ou échantillons) sont disposés en colonnes.

Caractères en colonnes : activez cette option si les caractères sont disposés en colonnes et les individus (ou échantillons) sont disposés en lignes.

X / Variables explicatives : sélectionnez une ou plusieurs variables explicatives qualitatives. Les données sélectionnées peuvent être de tout type, mais les données numériques sont automatiquement considérées comme nominales. Si le libellé des variables a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des variables : activez cette option si la première ligne des données sélectionnées (variables explicatives, libellés des observations) contient un libellé.

Libellés des observations : activez cette option si vous voulez utiliser des libellés d'observations pour l'affichage des résultats. Si l'option « Libellés des variables » est activée, la première cellule de la sélection doit comprendre un en-tête. Si vous n'activez pas cette option, des libellés seront automatiquement créés (Obs1, Obs2, …).

Onglet **Options** :

Type de test :

Paramétrique : activez cette option si vous souhaitez mettre en jeu des tests paramétriques (Student ou ANOVA à un facteur)

Non-paramétrique : activez cette option si vous souhaitez mettre en jeu le test non-paramétrique de Kruskal-Wallis. Cette option ne peut être activée que si une unique variable explicative est sélectionnée, et que cette variable soit qualitative.

Bayes empirique : activez cette option si vous souhaitez mettre en jeu le test bayésien empirique.

Niveau de signification (%) : entrez le niveau de signification à utiliser pour les différents tests (valeur par défaut : 5%).

Corrections post-hoc : sélectionnez le type de correction des p-values souhaité (Benjamini-Hochberg, Benjamini-Yekutieli, Bonferroni, pas de correction ; voir Description)

p-values à conserver : entrez le nombre de p-values les plus faibles à afficher dans les résultats. Si le nombre entré est supérieur au nombre de caractères du tableau caractères/individus filtré, XLSTAT affichera les p-values associées à l'intégralité des caractères. Valeur par défaut : 100.

Comparaisons multiples par paire : activez cette option puis choisissez la méthode de comparaison si vous le souhaitez. Des informations sur les tests de comparaisons multiples sont disponibles dans la section description.

Corrections de Bonferroni : activez cette option si vous souhaitez pénaliser les corrections multiples par paires par la méthode de Bonferroni.

Filtrage non spécifique : Activez cette option pour éliminer les caractères peu variables avant les calculs.

Critère et **seuil** : sélectionnez le critère de filtrage non-spécifique.

- Écart-type < : tous les caractères associés à un écart type inférieur au seuil choisi sont éliminés.
- Écart interquartile < : tous les caractères associés à un écart interquartile inférieur au seuil choisi sont éliminés.
- %(Écart-type) : un pourcentage de caractère associés à un faible écart-type sont éliminés. Ce pourcentage doit être indiqué dans la case "seuil".
- %(IQR) : un pourcentage de caractères associés à des écarts interquartiles faibles sont éliminés.

Onglet **Données manquantes** :

Ne pas accepter les valeurs manquantes : activez cette option pour que XLSTAT empêche la poursuite des calculs si des valeurs manquantes sont détectées.

Supprimer les observations : activez cette option pour supprimer les observations comportant des données manquantes.

Estimer les données manquantes : activez cette option pour estimer les données manquantes avant le début des calculs.

- **Moyenne ou mode** : activez cette option pour estimer les données manquantes en utilisant la moyenne (variables quantitatives) ou le mode (variables qualitatives) pour les variables correspondantes.
- **Plus proche voisin** : activez cette option pour estimer les données manquantes d'une observation en recherchant le plus proche voisin de l'observation.

Onglet **Sorties** :

Statistiques descriptives : activez cette option pour afficher les statistiques descriptives pour les variables sélectionnées.

Onglet **Graphiques** :

Histogramme des p-values : activez cette option pour afficher l'histogramme des p-values corrigées.

Volcano plot : activez cette option pour afficher le volcano plot (uniquement possible pour les variables explicatives à deux modalités).

Résultats

Statistiques descriptives : dans ce tableau sont affichées les statistiques descriptives correspondant aux différents caractères.

XLSTAT fournit les résultats suivants pour chaque variable explicative :

Tableau des x caractères associés à des p-values faibles : il contient de l'information sur les caractères les plus significatifs. Ceux-ci sont rangés par ordre croissant de p-value. La colonne p-values contient les p-values modifiées selon la méthode de correction post hoc sélectionnée. La colonne significative indique si la p-value concernée est significative par rapport au seuil alpha. Si l'option comparaisons multiples par paires a été activée, des colonnes supplémentaires s'affichent. Selon le type de test sélectionné, elles contiennent les moyennes (test paramétrique) ou médianes (test non- paramétrique) des modalités de la variable explicative. Au sein de chaque caractère, les modalités sont associées à des lettres résumant les résultats issus des comparaisons multiples. Deux modalités ne comprenant pas de lettre en commun sont significativement différentes. Deux modalités partageant une même lettre ne sont pas significativement différentes.

Graphiques : un histogramme représentant la distribution des p-values corrigées est suivi par un volcano plot permettant de repérer les caractères les plus intéressants en termes d'effets biologique et statistique.

Exemple

Un exemple d'étude d'expression différentielle est disponible sur le Centre d'aide XLSTAT à l'adresse suivante :

http://www.xlstat.com/demo-omicsdifff.htm

Bibliographie

Benjamini Y. and Hochberg Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, **57**, 289-300.

Benjamini Y. and Yekutieli D. (2001). The control of the false discovery rate in multiple hypothesis testing under dependency. *Annals of Statistics*, **29**, 1165-88.

Hahne F., Huber W., Gentleman R. and Falcon S. (2008). Bioconductor Case Studies. Springer.

Heat maps

Utilisez ce module pour effectuer une classification simultanée sur les lignes et les colonnes d'un tableau de données caractères/individus et générer des heat maps.

Dans cette section :

Description

Dans le cadre de l'exploration de tableaux de données caractères/individus, la heatmap permet de détecter d'éventuels groupements de caractères (gènes, protéines, métabolites) s'exprimant de manière similaire et caractérisant des groupements d'individus (échantillons) similaires. Par exemple, un groupement d'échantillons de rein malade peut être caractérisé par une forte expression d'un groupe particulier de gènes, en comparaison avec d'autres échantillons.

Construction d'une heat map dans XLSTAT

Les caractères et les individus sont classifiés de manière indépendante grâce à une classification ascendante hiérarchique centrée sur des distances euclidiennes, et précédée par une classification k-means (nuées dynamiques) si la matrice de données est très volumineuse. Les lignes et les colonnes de la matrice sont par la suite permutées par rapport à ces deux classifications, ce qui rapproche les lignes similaires les unes des autres et les colonnes similaires les unes des autres. Enfin, une heat map reflétant les données de la matrice est affichée : les valeurs sont remplacées par des intensités de couleur.

Filtrage non spécifique

Avant de lancer les analyses, il est intéressant d'éliminer les caractères dont l'expression est peu variable à travers les individus. Le filtrage non- spécifique a deux avantages principaux :

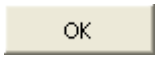
- Il fait en sorte que le calcul se focalise moins sur les caractères probablement non exprimés différentiellement.
- Il limite les pénalisations post-hoc, puisque le nombre de p-values calculées est plus faible.

Deux méthodes sont disponibles dans XLSTAT :

- L'utilisateur indique un seuil de variabilité (écart interquartile ou écart type). Les caractères dont la variabilité est plus faible que ce seuil sont éliminés en amont des analyses.
- L'utilisateur spécifie un pourcentage de caractères avec une faible variabilité (écart interquartile ou écart type) à éliminer en amont des analyses.

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.

Onglet **Général**:

Tableau des Caractères/Individus : sélectionnez ici la matrice de données caractères/individus. Les données sélectionnées doivent être du type numérique.

Format des données :

Caractères en lignes : activez cette option si les caractères sont disposés en lignes et les individus (ou échantillons) sont disposés en colonnes.

Caractères en colonnes : activez cette option si les caractères sont disposés en colonnes et les individus (ou échantillons) sont disposés en lignes.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés inclus : activez cette option si les libellés des caractères et ceux des individus sont compris dans la sélection.

Classer les caractères : activez cette option si vous souhaitez inclure une classification des caractères dans la heat map.

Classer les individus : activez cette option si vous souhaitez inclure une classification des individus (ou échantillons) dans la heat map.

Onglet **Options**:

Centrer : activez cette option pour center chaque ligne séparément.

Réduire : activez cette option pour réduire chaque ligne séparément.

Critère et **seuil** : sélectionnez le critère de filtrage non- spécifique.

- Écart-type< : tous les caractères associés à un écart type inférieur au seuil choisi sont éliminés.
- Écart interquartile< : tous les caractères associés à un écart interquartile inférieur au seuil choisi sont éliminés.
- %(Écart-type) : un pourcentage de caractère associés à un faible écart-type sont éliminés. Ce pourcentage doit être indiqué dans la case "seuil".
- %(IQR) : un pourcentage de caractères associés à des écarts interquartiles faibles sont éliminés.

Onglet **Données manquantes** :

Ne pas accepter les valeurs manquantes : activez cette option pour que XLSTAT empêche la poursuite des calculs si des valeurs manquantes sont détectées.

Supprimer les observations : activez cette option pour supprimer les observations comportant des données manquantes.

Estimer les données manquantes : activez cette option pour estimer les données manquantes avant le début des calculs.

- **Moyenne ou mode** : activez cette option pour estimer les données manquantes en utilisant la moyenne (variables quantitatives) ou le mode (variables qualitatives) pour les variables correspondantes.
- **Plus proche voisin** : activez cette option pour estimer les données manquantes d'une observation en recherchant le plus proche voisin de l'observation.

Onglet **Sorties**:

Statistiques descriptives : activez cette option pour afficher les statistiques descriptives des individus (échantillons).

Onglet **Graphiques** :

Echelle de couleur : choisissez la gamme de couleur de la heat map (rouge à vert via noir ; rouge à bleu via blanc ; rouge à jaune)

Calibration des couleurs :

- **Automatique** : Activez cette option pour qu'XLSTAT choisisse automatiquement les valeurs associées aux couleurs limites de la heat map.
- **Définie par l'utilisateur** : Activez cette option pour choisir manuellement les valeurs minimum (**Min**) et maximum (**Max**) associées aux couleurs limites de la heat map.

Largeur et **hauteur** : introduisez un facteur d'amplification pour la largeur ou la hauteur de la heat map.

Résultats

Statistiques descriptives : les tableaux de statistiques descriptives présentent les statistiques simples pour tous les individus. Le nombre d'observations, le nombre de données manquantes, le nombre de données non manquantes, la moyenne, et l'écart-type (non biaisé) sont affichés.

Heatmap : le dendrogramme des caractères est affiché verticalement (lignes) et le dendrogramme des individus est affiché horizontalement (colonnes). Au milieu, une heatmap reflétant les valeurs de données s'affiche.

Les groupements de caractères similaires sont caractérisés par des rectangles de couleur homogène traversant la heatmap horizontalement.

Les groupements d'individus similaires sont caractérisés par des rectangles de couleur homogène traversant la heatmap verticalement.

Les groupements d'individus similaires caractérisés par des groupements de caractères similaires apparaissent sous forme de rectangles ou carrés homogènes à l'intersection de groupements de caractères et de groupements d'individus.

Exemple

Un exemple de heap maps est disponible sur le Centre d'aide XLSTAT à l'adresse suivante :

http://www.xlstat.com/demo-omicsheatf.htm

Bibliographie

Hahne F., Huber W., Gentleman R. and Falcon S. (2008). Bioconductor Case Studies. Springer.

Tests d'aptitude interlaboratoires

Utilisez cet outil pour réaliser des tests d'aptitude pour un ou ou plusieurs partipants (laboratoires, organismes de contrôle, individus), lorsque qu'une ou plusieurs mesures (désignées par tests dans XLSTAT) ont été enregistrées pour chaque partipant.

Dans cette section :

Description

Les tests d'aptitude impliquent l'utilisation de méthodes statistiques pour comparer les performances de plusieurs participants (qui peuvent être des laboratoires, des organismes de contrôle ou des individus), appelés « partipants » dans XLSTAT, pour des mesures spécifiques (appelées « tests » dans XLSTAT). Les tests d'aptitude peuvent être effectués pour évaluer les performances des laboratoires effectuant des mesures, pour détecter des problèmes lorsqu'ils surviennent dans un ou plusieurs laboratoires, ou pour établir l'efficacité et la comparabilité de différentes méthodes.

Les méthodes consistent à identifier puis à supprimer ou à ignorer les valeurs aberrantes et à produire des estimations robustes des estimateurs de localisation et d'échelle.

Cet outil est basé sur la norme ISO-13528. Il a été développé à l'aide de l'édition 2015 de la norme. Certaines erreurs ayant été détectées dans la version 2015 du document (voir Fahmy, 2021 pour plus de détails), XLSTAT permet également aux utilisateurs d'effectuer l'analyse en incluant les erreurs.

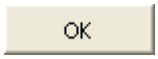
XLSTAT permet une analyse hautement automatisée des données de laboratoires et peut produire une série de statistiques classiques et robustes qui peuvent ensuite être utilisées pour interpréter les résultats et définir dans quelle mesure les normes sont respectées. Alors que de nombreuses fonctions XLSTAT peuvent être utilisées pour analyser des données inter-laboratoires, l'automatisation fournie ici est très utile lorsque l'utilisateur souhaite utiliser des algorithmes recommandés mais plus complexes, tels que l'algorithme A, l'algorithme S ou l'approche Q/Hampel.

La *technique du Cercle* (Van Nuland, 1992) est très efficace. Elle permet de comparer simultanément les moyennes et les variances de plusieurs partipants, tout en permettant d'identifier d'éventuelles valeurs aberrantes.

Cet outil est encore en évolution et vous êtes invités à soumettre tout commentaire ou demande.

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.

Onglet **Général** :

Tableau Partipants/Tests ou Tests/Partipants : sélectionnez les données correspondant aux données collectées pour chaque partipant avec un ou plusieurs tests enregistrés pour chaque partipant (généralement les partipants sont des laboratoires et les tests sont des mesures). Vous pouvez inclure dans la sélection les étiquettes des partipants et des tests. Dans ce cas, cochez les options correspondantes.

Format des données : * **Tableau Partipants/Tests** : choisissez cette option si les lignes correspondent aux partipants et les colonnes aux tests. * **Tableau Tests/Partipants** : choisissez cette option si les lignes correspondent aux tests et les colonnes aux partipants.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des partipants : activez cette option si les libellés des partipants sont inclus dans la sélection.

Libellés des tests : activez cette option si les libellés des tests sont inclus dans la sélection.

Onglet **Options** :

Position : choisissez si vous voulez utiliser la moyenne ou la médiane comme statistique de position.

Erreurs ISO-13528-2015 : activez cette option if you want XLSTAT to mimic the errors in the computation of the Q_n statistic.

Algorithme S :

- **Echelle** : choisissez si vous voulez utiliser l'amplitude ou l'écart-type comme statistique d'échelle.

Algorithme A :

- **Echelle** : choisissez si vous voulez utiliser l'amplitude, l'écart-type avec le test de Grubbs pour supprimer les valeurs aberrantes, le nIQR, le Q_n ou le Q comme statistique d'échelle.
- **Seulement si MAD=0**: activez cette option pour remplacer le MAD par la médiane des différences absolues avec la moyenne, si la MAD est nulle.
- **Mettre à jour $s^\star$** : Activez cette option si vous voulez mettre à jour l'estimateur robuste $s^\star$ à chaque itération.

Onglet **Données manquantes** :

Ne pas accepter les valeurs manquantes : activez cette option pour que XLSTAT empêche la poursuite des calculs si des valeurs manquantes sont détectées.

Supprimer les observations : activez cette option pour supprimer les observations contenant des données manquantes.

Ignorer: activez cette option pour que les données manquantes ne soient supprimées que lorsque cela est nécessaire.

Onglet **Sorties** :

Statistiques descriptives : activez cette option pour afficher les statistiques descriptives.

Scores Z : Activez cette option pour afficher les Z-scores. Plusieurs méthodes sont disponibles pour calculer les scores Z: * **Moyenne/Ecart-type** : c'est le moyen habituel pour calculer les scores Z. Le centrage est effectué en utilisant la moyenne et la standardisation en utilisant l'écart-type. * $\mathbf{m^\star/s^\star}$: Le centrage est effectué en utilisant les statistiques robuste $m^\star$ et $s^\star$. Notez quand dans ce cas, la formule pour la standardisation utilise au dénominateur $échelle = 1.25\frac{s^\star}{\sqrt{r}}$ où r est le nombre de répétitions. * **Référence/Ecart-type**: Le centrage est effectué en utilisant la valeur entrée par l'utilisateur et la standardisation en utilisant l'écart-type. * **Référence/s***: Le centrage est effectué en utilisant la valeur entrée par l'utilisateur et la standardisation en utilisant la statistique robuste $s^\star$. Notez que la formule au dénominateur is $échelle = 1.25\frac{s^\star}{\sqrt{r}}$ où r est le nombre de répétitions.

Onglet **Graphiques** :

Graphique d'homocédasticité : activez cette option pour afficher le graphique "cercle" avec les moyennes en abscisse et les écarts-types en ordonnée. * **Libellés des participants** : activez cette option pour étiqueter les points en utilisant les noms des participants. * Pour la mesure d'échelle (scale), deux options sont disponibles. Vous pouvez choisir de tracer les **écarts-types** ou les **amplitudes** en fonction des moyennes.

Graphique de contrôle des scores Z : activez cette option pour afficher le graphique de contrôle des scores Z.

Résultats

XLSTAT affiche plusieurs tableaux et un graphique pour aider à analyser et interpréter les résultats.

S'il y a plusieurs partipants (laboratoires) et si deux ou plusieurs tests (mesures) ont été effectués pour chacun d'eux, XLSTAT affiche une liste de statistiques récapitulatives pour chacun d'eux, y compris des statistiques robustes.

Les statistiques récapitulatives sont ensuite calculées pour l'ensemble des partipants.

Si cela est possible, le graphique d'homoscédasticité est affiché pour comparer la position et l'échelle des différents éléments en utilisant la technique du Cercle. Des lignes de confiance (90 %, 95 %, 99 %) sont affichées pour identifier les partipants présentant des valeurs potentiellement aberrantes.

Enfin, s'il y a plusieurs partipants (laboratoires) et si deux ou plusieurs tests (mesures) ont été effectués, XLSTAT affiche les résultats des algorithmes avancés décrits dans la norme ISO-13528 (Algorithme A, Algorithme S, Q/Hampel) qui sont conçus pour calculer les estimateurs de position et d'échelle de manière itérative. L'algorithme A et la méthode Q/Hampel sont utilisés pour obtenir des estimateurs robustes d'échelle et de localisation. L'algorithme S est utilisé pour estimer le paramètre d'échelle à partir d'écarts-types ou d'amplitudes.

Exemple

Un tutoriel expliquant comment utiliser les tests d'aptitude interlaboratoires est disponible sur le Centre d'aide XLSTAT :

[http://www.xlstat.com/demo-ilbf.htm]6

Bibliographie

Addinsoft (2021). d_n constants for n between 2 and 100. Addinsoft. https://xlst.at/iso-13528-en

Croux C. and Rousseeuw P. J. (1992). Time-efficient algorithms for two highly robust estimators of scale. In Proceedings of the 10th Symposium on Computational Statistics, Yadolah Dodge and Joe Whittaker (Eds.), Vol. 1., Springer-Verlag, Heidelberg, 411-428.

International Standards Organisation (2015). ISO 13528:2015(E), Statistical methods for use in proficiency testing by interlaboratory comparison. Second edition 2015-08-01. International

Standards Organisation, Geneva, Switzerland.

Rousseeouw P. J. and Croux C. (1993). Alternatives to the Median Absolute Deviation. *Journal of the American Statistical Association*, 88(424), 1273–1283.

Van Nuland Y. (1992). ISO 9002 and the circle technique. *Quality Engineering*, *5(2)*, 269-291.

Analyse de données multiblocs

Analyse Canonique des Corrélations

Utilisez l'Analyse Canonique des Corrélations (aussi dénommée analyse des corrélations canoniques ou CCorA), pour étudier la corrélation entre deux tableaux de données et pour extraire de ces tableaux un ensemble de variables canoniques telles que ces dernières soient le plus corrélées possible avec les deux tableaux et orthogonales entre elles.

Dans cette section :

Description

L'analyse canonique des corrélations (CCorA, aussi dénommée analyse des corrélations canoniques) est l'une des méthodes permettant d'étudier les relations entre deux tableaux de données. Découverte par Hotelling (1936) cette méthode a été très utilisée en écologie mais elle est depuis supplantée par la RDA (Analyse de Redondance) et par l'ACC (Analyse Canonique des Correspondances).

Contrairement à la RDA, cette méthode est symétrique et n'a donc pas pour but de créer des facteurs susceptibles de prédire les variables d'un tableau Y à partir des variables d'un tableau X. Etant donnés deux tableaux Y_1 et Y_2, La CCorA a pour but d'obtenir des vecteurs a_i et b_i tels que

$$\rho(i) = \frac{corr(Y_1 a_i, Y_2 b_i)}{var(Y_1 a_i) var(Y_2 b_i)}$$

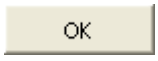
soit maximisé. Des contraintes doivent être introduites afin que la solution pour a_i et b_i soit unique. Comme on cherche finalement à maximiser la covariance entre $Y_1 a_i$ et $Y_2 b_i$ et à minimiser leur variance respective, il est possible d'obtenir des composantes bien corrélées entre elles, mais finalement peu représentatives des tableaux Y_1 et Y_2. Une fois la solution obtenue pour i=1, on cherche la solution pour i=2 où a_2 et b_2 doivent être respectivement orthogonaux à a_1 et b_2, et ainsi de suite. Le nombre de vecteurs que l'on peut obtenir est au maximum égal à $min(p, q)$ où p est le nombre de variables de Y_1 et q le nombre de variables de Y_2.

L'analyse inter-batteries de Tucker (1958) est une alternative où l'on cherche à maximiser uniquement la covariance entre les composantes Y1a(i) et Y2b(i).

L'analyse inter-batteries de Tucker (1958) est une alternative où l'on cherche à maximiser uniquement la covariance entre les composantes Y1a(i) et Y2b(i).

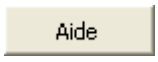
Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.



: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas (mode colonnes), XLSTAT considère que les sites sont en lignes et les objets/variables en colonnes. Si la flèche est vers la droite (mode lignes), XLSTAT considère que les objets/variables sont en lignes et les sites en colonnes.

Onglet **Général**:

Y1 : sélectionnez les données correspondant au premier tableau. Si des en-têtes de colonnes ont été sélectionnés (mode colonnes), veuillez vérifier que l'option « Libellés des colonnes » est activée. Si des en-têtes de lignes ont été sélectionnés (mode lignes), veuillez vérifier que l'option « Libellés des lignes » est activée.

Y2 : sélectionnez les données correspondant au second tableau. Si des en- têtes de colonnes ont été sélectionnés (mode colonnes), veuillez vérifier que l'option « Libellés des colonnes » est activée. Si des en-têtes de lignes ont été sélectionnés (mode lignes), veuillez vérifier que l'option « Libellés des lignes » est activée.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des colonnes/lignes : activez cette option si, en mode colonnes, la première ligne des données sélectionnées contient un libellé, ou si en mode lignes, la première colonne des données sélectionnées contient un libellé.

Libellés des observations : activez cette option si vous voulez utiliser des libellés les observations pour l'affichage des résultats. Si l'option « Libellés des colonnes » est activée (modes colonnes), la première cellule de la sélection doit comprendre un en-tête. Si vous n'activez pas cette option, des libellés seront automatiquement créés (Obs1, Obs2, …).

Onglet **Options**:

Type d'analyse : choisissez à partir de quel type de matrice de similarité doivent être calculées les corrélations canoniques.

Y1 :

- **Centrer** : activez cette option si vous voulez centrer les variables du tableau Y1.
- **Réduire** : activez cette option si vous voulez réduire les variables du tableau Y1.

Y2 :

- **Centrer** : activez cette option si vous voulez centrer les variables du tableau Y2.
- **Réduire** : activez cette option si vous voulez réduire les variables du tableau Y2.

Remarque : si les deux tableaux sont centrés-réduits, choisir le type d'analyse covariance ou corrélations donne le même résultat.

Filtrer les facteurs : vous pouvez activer l'une ou les deux options suivantes afin de réduire le nombre de facteurs pour lesquels les résultats sont affichés :

- **% minimum** : activez cette option puis saisissez le pourcentage minimum de la variabilité totale que doivent représenter les facteurs retenus.
- **Nombre maximum** : activez cette option pour fixer le nombre maximum de facteurs à prendre en compte.

Onglet **Données manquantes** :

Ne pas accepter les valeurs manquantes : activez cette option pour que XLSTAT empêche la poursuite des calculs si des valeurs manquantes sont détectées.

Supprimer les observations : activez cette option pour supprimer les observations comportant des données manquantes.

Estimer les données manquantes : activez cette option pour estimer les données manquantes avant le début des calculs.

- **Moyenne ou mode** : activez cette option pour estimer les données manquantes en utilisant la moyenne (variables quantitatives) ou le mode (variables qualitatives) pour les variables correspondantes.
- **Plus proche voisin** : activez cette option pour estimer les données manquantes d'une observation en recherchant le plus proche voisin de l'observation.

Onglet **Sorties**:

Statistiques descriptives : activez cette option pour afficher les statistiques descriptives pour les variables sélectionnées.

Covariance/Corrélations /[Y1Y2]'[Y1Y2] : activez cette option pour afficher la matrice de similarité utilisée.

Valeurs propres : activez cette option pour afficher le tableau et le graphique (scree plot) des valeurs propres.

Test du Lambda de Wilks : activez cette option pour afficher les résultats du test du Lambda de Wilks.

Corrélations canoniques : activez cette option pour afficher les corrélations canoniques. Ces dernières, comprises entre 0 et 1 seront d'autant plus élevées que la corrélation entre Y1 et Y2 est élevée.

Coefficients de redondance : activez cette option pour afficher les coefficients de redondance.

Coefficients canoniques : activez cette option pour afficher les coefficients canoniques. Ils correspondent aux coefficients associés à chacune des variables initiales pour la construction des variables canoniques. Ils sont standardisés si les variables initiales sont centrées réduites.

Corrélations Variables/Facteurs : activez cette option pour afficher les corrélations entre les variables initiales et les variables canoniques.

Coefficients d'adéquation des variables canoniques : activez cette option pour afficher les coefficients d'adéquation des variables canoniques.

Cosinus carrés : activez cette option pour afficher les cosinus carrés des variables initiales dans l'espace des variables canoniques.

Scores : activez cette option pour afficher les coordonnées des observations dans l'espace des variables canoniques.

Onglet **Graphiques** :

Graphiques de corrélations : activez cette option pour afficher les graphiques mettant en jeu des corrélations entre des composantes et des variables initiales.

- **Vecteurs** : activez cette option pour afficher les variables d'origine sous forme de vecteurs.
- **Etiquettes colorées** : activez cette option pour que les étiquettes soient de la même couleur que les points correspondants.

Résultats

Statistiques simples : le tableau de statistiques descriptives présente pour les deux tableaux sélectionnés des statistiques simples.

Matrice de similarité : la matrice utilisée pour les calculs et correspondant au choix fait dans la boîte de dialogue dans l'onglet « Options » est affichée.

Valeurs propres et pourcentages d'inertie : dans ce tableau sont affichés les valeurs propres, l'inertie associée, et les pourcentages de variabilité associés à chacune des variables canoniques. Remarque : dans d'autres logiciels, les valeurs propres fournies sont égales à L / (1-L), où L est la valeur propre fournie par XLSTAT.

Test du Lambda de Wilks : le test du Lambda de Wilks permet de déterminer si les deux tableaux Y1 et Y2 sont significativement liés à chacune des variables canoniques.

Corrélations canoniques : les corrélations canoniques, comprises entre 0 et 1, sont d'autant plus élevées que la corrélation entre Y1 et Y2 est élevée. Elles n'indiquent cependant pas à quel point les variables canoniques sont représentatives ou non de Y1 et Y2. Le carré d'une corrélation canonique est égal aux valeurs propres, et correspond donc au pourcentage de variabilité représenté par la variable canonique en question.

Les résultats ci-dessous sont calculés séparément pour chacun des deux groupes de variables initiales.

Coefficients de redondance : ces coefficients permettent pour chacun des deux tableaux de mesurer quel proportion de la variabilité des variables initiales est prédite par chacune des variables canoniques.

Coefficients canoniques : ces coefficients (en anglais *Canonical weights* , ou *Canonical function coefficients* ou *Canonical coefficients*) indiquent comment sont construites les variables canoniques, puisqu'ils correspondent aux coefficients de la combinaison linéaire qui permet de construire les variables canoniques à partir des variables initiales. Ils sont standardisés si les variables initiales sont centrées réduites. Dans ce cas, les poids relatifs des variables peuvent être comparés.

Les **corrélations entre les variables initiales et les variables canoniques** (appelées en anglais parfois *Structure correlation coefficients,* ou *Canonical factor loadings*). Elles permettent d'interpréter les variables canoniques.

Coefficients d'adéquation des variables canoniques : ces coefficients correspondent pour une variable canonique à la somme quadratique des corrélations entre variables initiales et variables canoniques, divisée par le nombre de variables initiales. Ils donnent le pourcentage de variabilité pris en compte par la variable canonique en question.

Cosinus carrés : les cosinus carrés des variables initiales dans l'espace des variables canoniques (qui correspondent aux carrés des corrélations entre variables initiales et variables canoniques), permettent de savoir si une variable initiale est bien représentée ou non dans l'espace des variables canoniques. La somme des cosinus carrés pour une variable initiale donnée est égale à 1 pour l'ensemble des variables canoniques. Lorsque l'on calcule cette somme pour un nombre réduit d'axes on parle de communalité (comme en analyse factorielle des variables latentes).

Scores : les scores correspondent aux coordonnées des observations dans l'espace des variables canoniques.

Exemple

Un exemple d'Analyse Canonique des Corrélations est disponible sur le Centre d'aide XLSTAT :

http://www.xlstat.com/demo-ccoraf.htm

Bibliographie

Hotelling H. (1936). Relations between two sets of variables. *Biometrika*, **28**, 321-327.

Jobson J.D. (1992). Applied Multivariate Data Analysis. Volume II: Categorical and Multivariate Methods. Springer-Verlag, New York.

Legendre P. and Legendre L. (1998). Numerical Ecology. Second English Edition. Elsevier, Amsterdam.

Tucker L.R. (1958). An inter-battery method of factor analysis. *Psychometrika*, **23(2)**,111-136.

Analyse de Redondance (RDA)

Utilisez l'Analyse de Redondance (*Redundancy Analysis* ou RDA en anglais), aussi appelée Analyse en Composantes Principales sur Variables Instrumentales (ACPVI), pour analyser un tableau de variables réponse tout en tenant compte de l'information fournie par des variables explicatives, et pour visualiser sur la même graphique les deux ensembles de variables, et les observations.

Dans cette section :

Description

L'Analyse de Redondance (RDA) a été développée par Van den Wollenberg (1977) comme alternative à l'Analyse Canonique des Corrélations (CCorA). La RDA permet d'étudier la relation entre deux tableaux de variables Y et X. Tandis que la CCorA est une méthode symétrique, la RDA est dissymétrique. Avec la CCorA, les composantes extraites des deux tableaux sont telles que leur corrélation est maximisée. Avec la RDA, les composantes extraites à partir de X sont telles qu'elles sont autant que possible corrélés avec les variables de Y. Les composantes de Y sont ensuite extraites de telle sorte qu'elles soient autant que possible corrélées avec les composantes extraites de X.

Principe de la RDA

Soit Y un tableau de variables réponse comprenant n observations et p variables. Ce tableau peut être analysé avec une analyse en composantes principales, afin d'obtenir une visualisation simultanée (biplot) des observations et des variables en deux ou trois dimensions.

Soit X un second tableau correspondant aux mesures pour les mêmes n observations de q variables quantitatives et/ou qualitatives.

L' analyse de redondance permet d'analyser la relation entre Y et X, et d'obtenir une représentation simultanée des observations, des variables réponse, et des variables explicatives en deux ou trois dimensions, optimale pour un critère de covariance (Ter Braak 1986).

L'analyse de redondance peut être décomposée en deux sous-parties :

une analyse sous contraintes dans un espace de dimension $min(n-1, p, q)$. Cette partie est celle qui présente le plus d'intérêt car elle permet de relier l'analyse du tableau Y à X. Cette analyse est dénommée RDA contrainte

une analyse de la partie résiduelle, non contrainte, dans un espace de dimension $min(n-1, p)$. Cette analyse est dénommée RDA non-contrainte.

RDA partielle

La RDA partielle ajoute une étape préliminaire. Le tableau X est subdivisé en deux groupes de variables : $X(1)$ comprend des variables de conditionnement dont on veut supprimer l'effet, déjà connu ou sans intérêt pour l'étude. Des régressions de Y et $X(2)$ par $X(1)$ sont calculés, et les résidus de ces régressions sont ensuite utilisés pour la RDA. La RDA partielle permet donc d'étudier l'effet du second groupe de variables, sans que les variables du premier groupe ne viennent perturber l'analyse.

La terminologie Observations/Variables réponse/Variables explicatives a été choisie dans XLSTAT. Dans le cadre d'une étude en écologie, « Sites » pourrait être utilisé à la place d' « Observations », « Espèces » à la place de « Variables réponse », et « Variables environnementales » à la place de « Variables explicatives ».

Problématique des facteurs de mise à l'échelle (scaling) pour les biplots

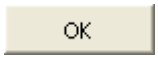
XLSTAT propose trois types de mise à l'échelle. Le type de mise à l'échelle change la façon sont calculées dont les coordonnées (aussi appelés scores) des variables réponse et des observations, ce qui modifie par conséquent, leur position respective sur la représentation graphique. Soit $u(ik)$ la coordonnée normalisée de la variable réponse i sur l'axe k, $v(ik)$ la coordonnée normalisée de l'observation i sur l'axe k, $L(k)$ la valeur propre correspondant à l'axe k, et T l'inertie totale (la somme des $L(k)$ pour les RDA contrainte et non-contrainte). Les trois mises à l'échelle proposées dans XLSTAT, identiques à celles de vegan (un module pour le logiciel de R, Oksanen, 2007). Les $u(ik)$ sont multipliés par c, et les $v(ik)$ par d, et r est une constante définie par $r = \sqrt[4]{(n-1)T}$, où n est le nombre d'observations.

$$\begin{aligned}
&\text{Scaling 1:} \quad c = r\sqrt{L(k)/T} \qquad d = r \\
&\text{Scaling 2:} \quad c = r \qquad d = r\sqrt{L(k)/T} \\
&\text{Scaling 3:} \quad c = r\sqrt[4]{L(k)/T} \qquad d = r\sqrt[4]{L(k)/T}
\end{aligned}$$

En plus des observations et des variables réponse, les variables explicatives peuvent être affichées sur le graphique. Les coordonnées de ces dernières sont obtenues en calculant les corrélations entre les variables du tableau X et les coordonnées des observations.

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas (mode colonnes), XLSTAT considère que les sites sont en lignes et les objets/variables en colonnes. Si la flèche est vers la droite (mode lignes), XLSTAT considère que les objets/variables sont en lignes et les sites en colonnes.

Onglet **Général**:

Variables réponse Y : sélectionnez le tableau correspond aux variables réponse. Si des en-têtes de colonnes ont été sélectionnés (mode colonnes), veuillez vérifier que l'option « Libellés des colonnes » est activée. Si des en-têtes de lignes ont été sélectionnés (mode lignes), veuillez vérifier que l'option « Libellés des lignes » est activée.

Variables explicatives X : sélectionnez le tableau correspondant aux variables explicatives mesurées pour les mêmes observations que Y.

- **Quantitatives** : activez cette option si vous disposez de variables quantitatives.
- **Qualitatives** : activez cette option si vous disposez de variables qualitatives.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

RDA partielle : activez cette option pour réaliser une RDA partielle. Si vous activez cette option une boîte de dialogue sera affichée au cours des calculs afin de vous permettre de sélectionner

quelles variables sont des variables de conditionnement (voir la section description).

Libellés des colonnes/lignes : activez cette option si, en mode colonnes, la première ligne des données sélectionnées contient un libellé, ou si en mode lignes, la première colonne des données sélectionnées contient un libellé.

Libellés des observations : activez cette option si vous voulez utiliser des libellés les observations pour l'affichage des résultats. Si l'option « Libellés des colonnes » est activée (modes colonnes), la première cellule de la sélection doit comprendre un en-tête. Si vous n'activez pas cette option, des libellés seront automatiquement créés (Obs1, Obs2, …).

Onglet **Options**:

Filtrer les facteurs : vous pouvez activer l'une ou les deux options suivantes afin de réduire le nombre de facteurs pour lesquels les résultats sont affichés :

- **% minimum** : activez cette option puis saisissez le pourcentage minimum de la variabilité totale que doivent représenter les facteurs retenus.
- **Nombre maximum** : activez cette option pour fixer le nombre maximum de facteurs à prendre en compte.

Test de permutation : activez cette option si vous voulez utiliser un test de permutation pour établir s'il existe ou non une relation entre les deux tableaux.

- **Nombre de permutations** : entrez le nombre de permutations à réaliser pour les tests (valeur par défaut : 300)
- **Niveau de signification (%)** : entrez le niveau de signification pour les tests.

Variables réponse :

- **Centrer** : activez cette option si vous voulez centrer les variables réponse avant de lancer la RDA.
- **Réduire** : activez cette option si vous voulez réduire les variables réponse avant de lancer la RDA.

Variables explicatives :

- **Centrer** : activez cette option si vous voulez centrer les variables explicatives avant de lancer la RDA.
- **Réduire** : activez cette option si vous voulez réduire les variables explicatives de lancer la RDA.

Type de biplot : choisissez le type de biplot à afficher. Les coordonnées (scores) des variables réponse et des observations sont calculées différemment en fonction du type choisi (voir la section description pour plus de détails).

Onglet **Données manquantes** :

Ne pas accepter les valeurs manquantes : activez cette option pour que XLSTAT empêche la poursuite des calculs si des valeurs manquantes sont détectées.

Supprimer les observations : activez cette option pour supprimer les observations comportant des données manquantes.

Estimer les données manquantes : activez cette option pour estimer les données manquantes avant le début des calculs.

- **Moyenne ou mode** : activez cette option pour estimer les données manquantes en utilisant la moyenne (variables quantitatives) ou le mode (variables qualitatives) pour les variables correspondantes.
- **Plus proche voisin** : activez cette option pour estimer les données manquantes d'une observation en recherchant le plus proche voisin de l'observation.

Onglet **Sorties**:

Statistiques descriptives : activez cette option pour afficher les statistiques descriptives pour les variables sélectionnées.

Résultats de la RDA : activez cette option pour afficher les résultats de la RDA contrrainte.

Résultats de l'ACC non contrainte : activez cette option pour afficher les résultats de la RDA non contrainte.

Valeurs propres : activez cette option pour afficher le tableau et le graphique (scree plot) des valeurs propres.

Scores (Observations) : activez cette option pour afficher les coordonnées (scores) des observations.

Scores (Variables réponse) : activez cette option pour afficher les coordonnées (scores) des variables réponse.

- **WA scores** : activez cette option pour calculer et afficher les « Weighted Average » scores.
- **LC scores** : activez cette option pour calculer et afficher les « Linear Combinations » scores.

Contributions : activez cette option pour afficher les contributions des observations et des variables réponse aux axes factoriels.

Cosinus carrés : activez cette option pour afficher les cosinus carrés des observations et des variables réponse avec les axes factoriels.

Onglet **Graphiques** :

Choisissez l'information que vous voulez afficher sur le biplot/triplot :

- **Observations** : activez cette option pour afficher les observations sur le graphique.
- **Variables réponse** : activez cette option pour afficher les variables réponse sur le graphique.
- **Variables explicatives** : activez cette option pour afficher les variables explicatives sur le graphique.

Etiquettes : activez cette option pour afficher les étiquettes sur les graphiques.

- **Etiquettes colorées** : activez cette option pour que les étiquettes soient de la même couleur que les points correspondants.

Vecteurs : activez cette option pour afficher des vecteurs.

- **Facteur d'allongement** : activez cette option pour jouer sur la longueur des vecteurs affichés.

Résultats

Statistiques simples : le tableau de statistiques descriptives présente pour les deux tableaux sélectionnés des statistiques simples.

Valeurs propres et pourcentages d'inertie : dans ces tableaux sont affichés pour la RDA contrainte et la RDA non contrainte, les valeurs propres, l'inertie associée, et les pourcentages correspondant, soit en terme d'inertie contrainte (ou non-contrainte), soit en terme d'inertie totale.

Les coordonnées (ou scores) des observations, des variables réponse et explicatives sont ensuite affichées. Ces coordonnées sont utilisées pour le graphique (simple, biplot ou triplot).

Le graphique permettent de visualiser la relation entre les observations, les variables réponse et explicatives. Lorsque des variables qualitatives ont été utilisées, les modalités correspondantes apparaissent en rouge avec un cercle évidé sur les graphiques. La légende les présente comme « modalités » afin de les différencier des autres variables explicatives.

Exemple

Un exemple d'Analyse de Redondance est disponible sur le Centre d'aide XLSTAT :

http://www.xlstat.com/demo-rdaf.htm

Bibliographie

Legendre P. and Legendre L. (1998). Numerical Ecology. Second English Edition. Elsevier, Amsterdam.

Oksanen J., Kindt R., Legendre P. and O'Hara R.B. (2007). vegan: Community Ecology Package version 1.8-5. http://cran.r-project.org/.

Ter Braak, C. J. F. (1992). Permutation versus bootstrap significance tests in multiple regression and ANOVA. in K.-H. Jöckel, G. Rothe, and W. Sendler, Editors. Bootstrapping and Related Techniques. Springer Verlag, Berlin.

Van den Wollenberg, A.L. (1977). Redundancy analaysis. An alternative for canonical correlation analysis. *Psychometrika*, **42(2)**, 207-219.

Analyse Canonique des Correspondances (ACC)

Utilisez l'analyse canonique des correspondances (en anglais, Canonical Correspondence Analysis, ou CCA), aussi appelée Analyse Factorielle des Correspondances sur Variables Instrumentales (ACPVI), pour analyser un tableau de contingence (typiquement un tableau de comptages, croisant sites et espèces) tout en tenant compte de l'information fournie par des variables quantitatives ou qualitatives mesurées sur les mêmes sites.

Dans cette section :

Description

L'analyse canonique des correspondances (en anglais, *Canonical Correspondence Analysis*, ou *CCA*) a été développée dans le cadre d'applications en écologie (Ter Braak, 1986). Néanmoins, cette méthode dont le cadre conceptuel est bien défini, peut être utilisée dans d'autres domaines. Le géomarketing et les analyses démographiques devraient pouvoir en tirer profit.

Principe de l'ACC

Soit T1 un tableau de contingence correspondant au comptage en n sites des effectifs de p objets. Ce tableau peut être analysé avec une analyse factorielle des correspondances (AFC) afin d'obtenir une visualisation simultanée des sites et des objets en deux ou trois dimensions.

Soit T2 un tableau correspondant aux mesures en les mêmes n sites de q variables quantitatives et/ou qualitatives.

L'analyse canonique des correspondances permet d'analyser la relation entre T1 et T2, et d'obtenir une représentation simultanée des sites, des objets, et des variables en deux ou trois dimensions, optimale pour un critère de variance (Ter Braak 1986, Chessel 1987).

L'analyse canonique des correspondances peut être décomposée en deux parties :

une analyse sous contraintes dans un espace de dimension q. Cette partie est celle qui présente le plus d'intérêt car elle permet de relier l'analyse du tableau T1 à T2.

une analyse de la partie résiduelle, non contrainte, dans un espace de dimension min(n-1-q, p-1). Cette analyse est dénommée ACC non-contrainte.

ACC partielle

L'ACC partielle ajoute une étape préliminaire. Le tableau T2 est subdivisé en deux groupes de variables : le premier contient des variables de conditionnement dont on veut supprimer l'effet, déjà connu ou sans intérêt pour l'étude, en réalisant une première ACC ; le second contient les variables dont on veut étudier l'effet. Une ACC est alors réalisée sur le tableau des résidus de la première ACC. L'ACC partielle permet donc d'étudier l'effet du second groupe de variables, sans que les variables du premier groupe ne viennent perturber l'analyse.

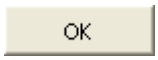
PLS-ACC

Tenenhaus (1998) a montré la possibilité d'utiliser la PLS dans le contexte de l'ACC. Addinsoft est le premier éditeur à proposer une intégration complète et efficace entre les deux méthodes. En utilisant une restructuration des données inspirée de la proposition de Tenenhaus, une étape PLS est appliquée aux données, soit pour créer des composantes PLS orthogonales optimales pour l'ACC qui permettent d'éviter les contraintes de l'ACC en termes de nombre de variables utilisables, soit pour sélectionner les variables les plus influentes avant de réaliser l'ACC. Les calculs de la seconde étape étant réalisés suivant la méthode classique d'ACC et les résultats habituels étant proposés, les utilisateurs coutumiers de l'ACC peuvent voir cette méthode comme une méthode de sélection de variables permettant de réduire le nombre de variables ou simplement de visualiser leur importance relative grâce au graphique des VIP (voir la section sur la régression PLS). Dans le cas d'une ACC partielle, l'étape préliminaire est inchangée.

La terminologie Sites/Objets/Variables a été choisie dans XLSTAT. « Individus » ou « observations » pourraient être utilisés à la place de « sites », et « espèces » pourrait être utilisé à la place de « objets » dans le cadre d'une étude en écologie.

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.

 : cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas (mode colonnes), XLSTAT considère que les sites sont en lignes et les objets/variables en colonnes. Si la flèche est vers la droite (mode lignes), XLSTAT considère que les objets/variables sont en lignes et les sites en colonnes.

Onglet **Général**:

Données Sites/Objets : sélectionnez le tableau de contingence correspondant aux comptages des différents objets en chacun des sites. Si des en-têtes de colonnes ont été sélectionnés (mode colonnes), veuillez vérifier que l'option « Libellés des colonnes » est activée. Si des en-têtes de lignes ont été sélectionnés (mode lignes), veuillez vérifier que l'option « Libellés des lignes » est activée.

Données Sites/Variables : sélectionnez le tableau correspondant aux différentes variables mesurées en chacun des sites.

- **Quantitatives** : activez cette option si vous disposez de variables quantitatives.
- **Qualitatives** : activez cette option si vous disposez de variables qualitatives.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

ACC partielle : activez cette option pour réaliser une ACC partielle. Si vous activez cette option une boîte de dialogue sera affichée au cours des calculs afin de vous permettre de sélectionner quelles variables sont des variables de conditionnement (voir la section description).

Libellés des colonnes/lignes : activez cette option si, en mode colonnes, la première ligne des données sélectionnées contient un libellé, ou si en mode lignes, la première colonne des données sélectionnées contient un libellé.

Libellés des sites : activez cette option si vous voulez utiliser des libellés des sites disponibles sur une feuille Excel pour l'affichage des résultats. Si l'option « Libellés des colonnes » est activée (modes colonnes), la première cellule de la sélection doit comprendre un en-tête. Si vous n'activez pas cette option, des libellés seront automatiquement créés (Obs1, Obs2, …).

ACC : activez cette option si vous voulez utiliser une ACC classique.

PLS-ACC : activez cette option si vous voulez utiliser une PLS-ACC (voir la section description).

Onglet **Options**:

Filtrer les facteurs : vous pouvez activer l'une ou les deux options suivantes afin de réduire le nombre de facteurs pour lesquels les résultats sont affichés :

- **% minimum** : activez cette option puis saisissez le pourcentage minimum de la variabilité totale que doivent représenter les facteurs retenus.
- **Nombre maximum** : activez cette option pour fixer le nombre maximum de facteurs à prendre en compte.

Test de permutation : activez cette option si vous voulez utiliser un test de permutation pour établir s'il existe ou non une relation entre les deux tableaux.

- **Nombre de permutations** : entrez le nombre de permutations à réaliser pour les tests (valeur par défaut : 300)
- **Niveau de signification (%)** : entrez le niveau de signification pour les tests.

PLS-ACC : si vous avez choisi l'option PLS-ACC, les options suivantes vous sont proposées :

- **Automatique** : choisissez cette option si vous voulez que XLSTAT détermine automatiquement par XLSTAT combien de composantes PLS doivent être gardées pour l'étape ACC.
- Définie par l'utilisateur :
- **Max composantes** : activez cette option pour fixer le nombre maximum de composantes à prendre en compte dans le modèle. La valeur par défaut est 2. Si cette option n'est pas activée, le nombre de composantes est déterminé automatiquement par XLSTAT.
- **Nombre de variables** : activez cette option pour définir le nombre de variables qui doivent être utilisées pour l'étape de l'ACC. Les variables qui ont les VIP les plus élevées sont sélectionnées. Les VIP sont celles du modèle avec le nombre de composantes retenues définies avec l'option "Max composantes".

Onglet **Données manquantes** :

Ne pas accepter les valeurs manquantes : activez cette option pour que XLSTAT empêche la poursuite des calculs si des valeurs manquantes sont détectées.

Supprimer les observations : activez cette option pour supprimer les observations comportant des données manquantes.

Estimer les données manquantes : activez cette option pour estimer les données manquantes avant le début des calculs.

- **Moyenne ou mode** : activez cette option pour estimer les données manquantes en utilisant la moyenne (variables quantitatives) ou le mode (variables qualitatives) pour les variables correspondantes.
- **Plus proche voisin** : activez cette option pour estimer les données manquantes d'une observation en recherchant le plus proche voisin de l'observation.

Onglet **Sorties**:

Statistiques descriptives : activez cette option pour afficher les statistiques descriptives pour les variables sélectionnées.

Profils lignes et colonnes : activez cette option pour afficher les profils lignes et les profils colonnes.

Résultats de l'ACC : activez cette option pour afficher les résultats de l'ACC.

Résultats de l'ACC non contrainte : activez cette option pour afficher les résultats de l'ACC non contrainte.

Valeurs propres : activez cette option pour afficher le tableau et le graphique (scree plot) des valeurs propres.

Coordonnées principales : activez cette option pour afficher les coordonnées principales des sites, des objets et des variables.

Coordonnées standard : activez cette option pour afficher les coordonnées standard des sites, des objets et des variables.

Contributions : activez cette option pour afficher les contributions.

Cosinus carrés : activez cette option pour afficher les cosinus carrés avec les axes factoriels.

Moyennes pondérées : activez cette option pour afficher les moyennes pondérées associées aux variables du tableau sites/variables.

Coefficients de régression : activez cette option pour afficher les coefficients de régression associés aux différentes variables dans l'espace factoriel.

Onglet **Graphiques** :

Sites et objets :

- **Sites et objets / Symétriques** : activez cette option pour afficher le graphique symétrique des sites et des objets. Les coordonnées principales sont utilisées pour les sites et les objets.

- **Sites / Asymétrique** : activez cette option pour afficher le graphique asymétrique des sites. Les coordonnées principales sont utilisées pour les sites, et les coordonnées standard pour les objets.

- **Objets / Asymétrique** : activez cette option pour afficher le graphique asymétrique des objets. Les coordonnées principales sont utilisées pour les objets, et les coordonnées standard pour les sites.

- **Sites** : activez cette option pour afficher un graphique sur lequel ne figurent que les sites. Les coordonnées principales sont utilisées.

- **Objets** : activez cette option pour afficher un graphique sur lequel ne figurent que les objets. Les coordonnées principales sont utilisées.

Variables :

- **Corrélations** : activez cette option pour afficher les variables quantitatives et qualitatives sur les graphiques, en utilisant comme coordonnées leurs corrélations (égales aux coordonnées standard).

- **Coefficients de régression** : activez cette option pour afficher les variables quantitatives et qualitatives sur les graphiques, en utilisant comme coordonnées les coefficients de régression correspondant.

Etiquettes : activez cette option pour afficher les étiquettes sur les graphiques.

- **Etiquettes colorées** : activez cette option pour que les étiquettes soient de la même couleur que les points correspondants.

Vecteurs : activez cette option pour afficher des vecteurs.

- **Facteur d'allongement** : activez cette option pour jouer sur la longueur des vecteurs affichés.

Résultats

Statistiques simples : le tableau de statistiques descriptives présente pour les deux tableaux sélectionnés des statistiques simples.

Inertie : dans ce tableau est affichée la répartition de l'inertie entre l'ACC contrainte et l'ACC non contrainte.

Valeurs propres et pourcentages d'inertie : dans ces tableaux sont affichés pour l'ACC contrainte et l'ACC non contrainte, les valeurs propres, l'inertie associée, et les pourcentages

correspondant, soit en terme d'inertie contrainte (ou non-contrainte), soit en terme d'inertie totale.

Moyennes pondérées : dans ce tableau sont affichées les moyennes pondérées pour chacun des sites, ainsi que les moyennes pondérées globales.

Pour l'ensemble les sites, les objets et les variables sont ensuite affichées les **coordonnées principales**, les **coordonnées standard**. Ces coordonnées sont utilisées pour les différents graphiques générés ensuite.

Coefficients de régression : dans ce tableau sont affichés les coefficients de régression des variables sur les axes factoriels.

Les graphiques permettent de visualiser la relation entre les sites, les objets et les variables. Lorsque des variables qualitatives ont été utilisées, les modalités correspondantes apparaissent en rouge avec un cercle évidé sur les graphiques. La légende les présente comme « modalités » afin de les différencier des autres variables explicatives.

Exemple

Un exemple d'Analyse Canonique des Correspondances est disponible sur le Centre d'aide XLSTAT :

http://www.xlstat.com/demo-ccaf.htm

Bibliographie

Chessel D., Lebreton J.D and Yoccoz N. (1987). Propriétés de l'analyse canonique des correspondances; une illustration en hydrobiologie. *Revue de Statistique Appliquée,* **35(4)**, 55-72.

Legendre P. and Legendre L. (1998). Numerical Ecology. Second English Edition. Elsevier, Amsterdam.

McCune B. (1997). Influence of noisy environmental data on canonical correspondence analysis. *Ecology*, **78(8)**, 2617-2623.

Palmer M.W. (1993). Putting things in even better order: The advantages of canonical correspondence analysis. *Ecology*, **74(8)**, 2215-2230.

Tenenhaus M. (1998). La Régression PLS, Théorie et Pratique. Technip, Paris.

Ter Braak C. J. F. (1986). Canonical Correspondence Analysis: a new eigenvector technique for multivariate direct gradient analysis. *Ecology*, **67(5)**, 1167-1179.

Ter Braak C. J. F. (1992). Permutation versus bootstrap significance tests in multiple regression and ANOVA. in K.-H. Jöckel, G. Rothe, and W. Sendler, Editors. Bootstrapping and Related Techniques. Springer Verlag, Berlin.

Analyse en Coordonnées Principales (PCoA)

Utilisez l'analyse en coordonnées principales (en anglais, *Principal Coordinate Analysis*) pour représenter graphiquement une matrice carrée décrivant la similarité ou la dissimilarité entre p éléments (individus, variables, objets, …).

Dans cette section :

Description

L'analyse en coordonnées principales (en anglais, *Principal Coordinate Analysis* ou *PCoA*) a pour but de représenter graphiquement une matrice de ressemblance entre p éléments (individus, variables, objets, …).

Si la matrice en entrée est une matrice de similarité, XLSTAT la transformera en une matrice de dissimilarité avant de faire les calculs proposés par Gower (1966) avec d'éventuelles modifications proposés par divers auteurs dont on trouve la synthèse dans le livre *Numerical Ecology* de Legendre et Legendre (1998).

Principe de l'analyse

Soit D la matrice $p \times p$ symétrique contenant les distances entre p éléments : on calcule alors une matrice A dont les éléments $a(ij)$ correspondant à la i-ème ligne et à la j-ème colonne sont définis par :

$$a(ij) = -d^2(ij)/2$$

On centre alors la matrice A par ligne et par colonne pour obtenir la matrice Δ_1 dont les élements $\delta_1(ij)$ sont donnés par :

$$\delta_1(ij) = a(ij) - \overline{a}(i) - \overline{a}(j) - \overline{a}$$

où $\overline{a}(i)$ est la moyenne des $a(ij)$ pour la ligne i, $\overline{a}(j)$ est la moyenne des $a(ij)$ pour la colonne j et $\overline{a}$ est la moyenne de tous les éléments.

On calcule alors la décomposition en valeurs propres de la matrice Δ_1. Les vecteurs propres sont triés par ordre décroissant de valeurs propres, et transformés de telle sorte que, si $u(k)$

est le vecteur propre associé à la valeur propre $\lambda(k)$, on ait :

$$u'(k)u(k) = \lambda(k)$$

Les vecteurs propres ainsi transformés sont les coordonnées principales, qui peuvent alors être directement utilisées pour représenter les p objets dans un espace à 1, 2, … $p-1$ dimensions.

Comme avec l'ACP (Analyse en Composantes Principales) les valeurs propres peuvent être interprétées en terme de pourcentage de variabilité représenté.

Remarque : parce que la matrice Δ_1 est centrée, on obtient au plus $p-1$ valeurs propres non nulles. Dans le cas ou la matrice de départ D est une matrice euclidienne, on comprend aisément que $p-1$ axes suffiront toujours à décrire p objets (par deux points passe une ligne, trois points sont toujours contenus dans un plan, …). Dans le cas où des points sont confondus dans un sous-espace, on peut obtenir plusieurs valeurs propres nulles (par exemple, trois points peuvent être alignés sur une même ligne).

Cas de valeurs propres négatives

Lorsque la matrice D n'est pas une matrice de distances métriques (cas de distances semi-métriques ou non métriques par exemple), ou si des valeurs manquantes étaient présentes dans les données ayant été utilisées pour calculer les distances, la décomposition en valeurs propres peut engendrer des valeurs propres négatives. Ce problème est décrit en détail dans l'article de Gower et Legendre (1986).

XLSTAT propose deux transformations pour remédier au problème des valeurs propres. La première consiste simplement à prendre la racine carrée des éléments de la matrice D. La seconde, inspirée de Lingoes (1971), consiste à ajouter une constante à la matrice D (sauf la diagonale qui reste nulle), telle qu'il n'y ait plus de valeurs propres négatives. Cette constante est égale à la valeur propre négative la plus élevée en valeur absolue.

Lorsqu'il y a des valeurs propres négatives, la représentativité des axes est calculée en appliquant la modification proposée par Caillez et Pagès (1976).

ACP, MDS et PCoA

L'ACP et la PCoA sont assez proches en ce sens que l'ACP permet aussi de représenter des individus dans un espace de faible dimension avec des axes optimaux en terme de variabilité représentée. La PCoA appliquée à la matrice des distances euclidiennes entre les individus (calculée après normalisation des colonnes avec l'écart-type non biaisé) aboutit au même résultat que l'ACP normée appliquée aux données brutes. Les valeurs propres issues de la PCoA sont égales à $(p-1)$ fois celles obtenues à partir de l'ACP.

La PCoA est une méthode dont le but est identique à celui du MDS (Multidimensional Scaling), à savoir représenter des objets pour lesquels on dispose d'une matrice de proximité.

Le MDS présente deux désavantages par rapport à la PCoA :

- l'algorithme est beaucoup plus complexe et plus lent ;

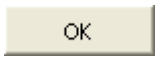
- les axes issus du MDS ne sont pas interprétables en terme de variabilité portée.

Le MDS présente deux avantages par rapport à la PCoA :

- l'algorithme s'accommode de données manquantes dans la matrice de proximité.
- la version non-métrique du MDS permet de traiter, sans que cela ne pose de problème théorique, des cas de matrices de proximité ou seul l'ordre compte.

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas (mode colonnes), XLSTAT considère que les sites sont en lignes et les objets/variables en colonnes. Si la flèche est vers la droite (mode lignes), XLSTAT considère que les objets/variables sont en lignes et les sites en colonnes.

Onglet **Général**:

Données : sélectionnez une matrice de similarité ou dissimilarité. Si seule la partie triangulaire inférieure ou supérieure est disponible, le tableau est accepté. Si des différences sont détectées entre les parties inférieure et supérieure de la matrice sélectionnée, XLSTAT vous en avertit, et vous propose de modifier les données (calcul de la moyenne des deux parties) pour pouvoir poursuivre les calculs.

Dissimilarités / Similarités : choisissez l'option correspondant à la nature des données de matrice sélectionnée.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés inclus : activez cette option si vous avez inclus les libellés des lignes et des colonnes dans la sélection.

Onglet **Options**:

Correction pour les valeurs propres négatives : activez l'une des options suivantes pour définir la stratégie à adopter si une valeur négative est détectée :

- **Aucune** : aucune correction n'est effectuée.
- **Racine carrée** : les éléments de la matrice des distances sont transformés en leur racine carrée.
- **Lingoes** : une transformation est effectuée pour supprimer la présence de valeur propre négative.

Filtrer les facteurs : vous pouvez activer l'une ou les deux options suivantes afin de réduire le nombre de facteurs pour lesquels les résultats sont affichés :

- **% minimum** : activez cette option puis saisissez le pourcentage minimum de la variabilité totale que doivent représenter les facteurs retenus.
- **Nombre maximum** : activez cette option pour fixer le nombre maximum de facteurs à prendre en compte.

Onglet **Sorties**:

Matrice Delta1 : activez cette option pour afficher la matrice Δ_1 utilisée pour le calcul des valeurs propres.

Valeurs propres : activez cette option pour afficher le tableau et le graphique (scree plot) des valeurs propres.

Coordonnées principales : activez cette option pour afficher les coordonnées principales des objets.

Contributions : activez cette option pour afficher le tableau des contributions.

Cosinus carrés : activez cette option pour afficher le tableau des cosinus carrés.

Onglet **Graphiques** :

Graphique : activez cette option pour afficher le graphique.

Résultats

Matrice Delta1 : cette matrice correspond à la matrice Δ_1 de Gower, utilisée pour le calcul des valeurs propres.

Valeurs propres et pourcentages d'inertie : dans ce tableau sont affichés les valeurs propres et les pourcentages de variabilité correspondant.

Coordonnées principales : dans ce tableau sont affichées les coordonnées principales des objets, utilisées pour le graphique qui permet d'interpréter les proximités entre ces derniers.

Contributions : utilisez ce tableau pour évaluer la contribution de chacun des objets à la construction de l'axe.

Cosinus carrés : utilisez ce tableau pour savoir à quel point un objet est proche d'un axe.

Exemple

Un exemple d'Analyse en Coordonnées Principales est disponible sur le Centre d'aide XLSTAT :

http://www.xlstat.com/demo-pcoaf.htm

Bibliographie

Cailliez F. and Pagès J.P. (1976). Introduction à l'Analyse des Données. Société de Mathématiques Appliquées et de Sciences Humaines, Paris.

Gower J. C. (1966). Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika*, **53**, 325-338.

Gower J.C. and Legendre P. (1986). Metric and Euclidean properties of dissimilarity coefficients. *Journal of Classification*, **3**, 5-48.

Legendre P. and Legendre L. (1998). Numerical Ecology. Second English Edition. Elsevier, Amsterdam.

Lingoes J.C. (1971). Some boundary conditions for a monotone analysis of symmetric matrices. *Psychometrika*, **36**, 195-203.

XLSTAT-PLSPM

XLSTAT-PLSPM est un module de XLSTAT qui permet d'appliquer des méthodes d'équations structurelles basées sur les composantes. Parmi celles-ci, on trouve l'approche PLS (moindres carrés partiels, *Partial Least Squares Path Modeling*), l'approche GSCA (Generalized Structured Components Analysis) et la méthode RGCCA (Regularized Generalized Canonical Correlation Analysis). Ces méthodes novatrices permettent de représenter simplement des relations complexes entre des variables observées et des variables non observées dites latentes.

XLSTAT-PLSPM comporte des méthodes ayant notamment des applications en marketing tel que l'analyse de la satisfaction des consommateurs. Trois niveaux d'affichage sont proposés afin de s'adapter aux différents utilisateurs de ces approches.

Dans cette section :

Description

L'approche PLS est une méthode statistique permettant de modéliser des relations complexes entre des variables observées et des variables latentes. Ce type de modèles est généralement appelé modèle d'équations structurelles à variables latentes. Depuis quelques années, cette approche est de plus en plus populaire dans des communautés scientifiques très variées (Esposito Vinzi et al., 2008). Les modèles d'équations structurelles (Structural Equation Models) comprennent un grand nombre de méthodologies statistiques (dont l'approche PLS fait partie) qui permettent l'estimation de relation de causalité complexes entre des variables latentes mesurées elles-mêmes par des variables observées dites manifestes.

L'approche PLS dans sa version actuelle a été présentée pour la première fois par Wold en 1979, mais les articles de référence sur cette méthode sont Wold (1982 et 1985).

Dans le cadre des modèles d'équations structurelles, deux méthodes s'opposent : d'une part, la méthode par analyse de la structure de covariance (bien souvent appelée LISREL) développée par Jöreskog (1970) et, d'autre part, l'approche PLS. Herman Wold a toujours opposé la première qui utilisait, selon ses termes, une « modélisation dure » (« hard modeling », hypothèses de distribution fortes, nécessité d'avoir plusieurs centaines d'observations) à la seconde basée sur une « modélisation douce » (« soft modeling », peu d'hypothèses de distribution et un très petit nombre d'observations suffit à son application). Les deux approches ont été comparées dans Jöreskog et Wold (1982).

Du point de vue classique des modèles d'équations structurelles à variables latentes, l'approche PLS est une méthode basée sur des composantes pour laquelle la causalité est formulée en termes d'espérance conditionnelle linéaire. L'approche PLS privilégie la recherche d'une optimalité prédictive des relations plutôt que celle de relation de causalité. Elle est orientée de manière prédictive afin de tester des hypothèses de causalité. Ainsi, plutôt que de valider un modèle en termes de qualité d'ajustement, on utilisera des indices de qualité prédictive que nous présentons par la suite. Pour plus de détails sur ces points, on peut voir deux articles de référence sur le sujet : Chin (1998, plutôt orienté vers les applications) et Tenenhaus et al. (2005, plutôt orienté vers la théorie).

Par ailleurs, l'approche PLS permet d'analyser des tableaux multiples et peut être directement reliée à des méthodes d'analyse de données classiques de ce domaine. En fait, l'approche PLS peut aussi être vue comme une méthode extrêmement flexible dans l'analyse de tableaux multiples grâce à, d'une part, l'approche PLS hiérarchique et, d'autre part, l'approche PLS confirmatoire (Tenenhaus et Hanafi, 2008). Ces approches montrent que les méthodes classiques basées sur les données (« data-driven methods ») peuvent être reliées à des méthodes basées sur la théorie (« theory-driven methods ») telle que les modèles d'équations structurelles). Ceci permet d'intégrer des connaissances sur les relations entre les tableaux dans les analyses.

D'autres méthodes ont été introduites plus récemment et sont disponibles dans XLSTAT-PLSPM. Notamment Generalized Structured Components Analysis (GSCA) ou Regularized Generalized Canonical Correlation Analysis (RGCCA).

L'algorithme de l'approche PLS

Un modèle structurel PLS est décrit par deux sous-modèles : (1) le modèle de mesure (ou modèle externe) reliant les variables manifestes (observées) aux variables latentes qui leur sont associées et (2) le modèle structurel (ou modèle interne) reliant des variables latentes dites endogènes à d'autres variables latentes.

1. Standardisation des variables manifestes

Il existe quatre options afin de standardiser les variables manifestes qui devront être choisies en fonction de certaines conditions sur les données initiales:

- Condition 1: Les échelles des variables manifestes sont comparables. Par exemple, dans l'exemple du tutoriel basé sur le modèle ECSI, les valeurs prises par les variables manifestes sont toutes entre 0 et 100 et sont donc comparables. D'autre part, on ne pourra pas comparer un poids en tonnes à une vitesse en km/h.

- Condition 2: Les moyennes des variables manifestes peuvent être interprétées. Par exemple, si la différence entre deux variables manifestes n'est pas analysable, alors les moyennes ne servent à rien.

- Condition 3: Les variances des variables manifestes traduisent l'importance de celles-ci.

- Si la condition 1 n'est pas vérifiée, alors il faut standardiser les variables manifestes (avec moyenne 0 et variance 1).

- Si la condition 1 est vérifiée, il peut être intéressant d'utiliser les informations venant des données. Mais l'estimation des paramètres dépend de la vérification des autres conditions :

- Les conditions 2 et 3 ne sont pas vérifiées : Les variables manifestes sont standardisées (avec moyenne 0 et variance 1) pour l'estimation des paramètres puis sont remises dans leur échelle originale afin d'obtenir l'estimation finale des poids et des loadings.

- La condition 2 est vérifiée mais la condition 3 ne l'est pas : Les variables manifestes ne sont pas centrées mais leurs variances sont standardisées à 1 pour la phase d'estimation des paramètres. Puis les variances des variables manifestes sont remises à leur valeur originale afin d'obtenir l'estimation finale des poids et des loadings.

- Les conditions 2 et 3 sont vérifiées : On utilise les variables manifestes originales.

Lohmöller (1989) a introduit un paramètre de standardisation afin de sélectionner l'une de ces quatre options :

Echelles des Variables comparables	Moyennes interprétables	Variances reliées à l'importance de la variable	Moyenne	Variance	Remise à l'échelle	METRIC
Non			0	1	Non	1
Oui	Non	Non	0	1	Oui	2
Oui	Oui	Non	Original	1	Oui	3
Oui	Oui	Oui	Original	Original		4

Avec METRIC=1, cas « standardisé, poids sur variables manifestes standardisées », METRIC=2, cas « standardisé, poids sur VM d'origine », METRIC=3, cas « réduit, poids sur VM d'origine », METRIC=4, cas « VM d'origine ».

2. Le modèle de mesure

Une variable latente (VL) ξ est une variable non observable (ou un construit) qui peut être décrit par un ensemble de variables observées x_h appelées variables manifestes (VM) ou indicateurs. Il y a trois manières de relier les variables manifestes à leur variable latente appelés respectivement la manière réflective, la manière formative et la manière MIMIC (Multiple effect Indicators for MultIple Causes).

2.1. La manière réflective

2.1.1. Définition

Dans le modèle, chaque variable manifeste est le reflet de la variable latente qui lui est associée. Chaque variable manifeste est reliée à sa variable latente par une simple équation de régression linéaire :

$$x_h = \pi_{h0} + \pi_h \xi + \epsilon_h \quad (1)$$

où ξ a pour moyenne m et pour écart-type 1. C'est un schéma réflectif: chaque variable manifeste est le reflet de la variable latente qui lui est associée. La seul hypothèse nécessaire dans le cas de ce modèle est que :

$$E(x_h|\xi) = \pi_{h_0} + \pi_h \xi \quad (2)$$

Cette hypothèse implique que le résidu ϵ_h a une moyenne de 0 et n'est pas corrélé à la variable latente ξ.

2.1.2. Vérification de l'unidimensionnalité des blocs

Dans le cas d'un modèle réflectif, les blocs de variables manifestes doivent être unidimensionnels au sens de l'analyse factorielle. Sur des données réelles, cette hypothèse doit être vérifiée. Trois outils principaux existent pour la vérifier : l'analyse en composantes

principales sur chaque bloc de variables manifestes, l'α de Cronbach et le ρ de Dillon-Goldstein.

a) Analyse en composantes principales d'un bloc

Un bloc est unidimensionnel lorsque la première valeur propre de la matrice de corrélation entre les variables manifestes du bloc est plus grande que 1 et la seconde est plus petite que 1 ou tout du moins beaucoup plus petite que la première. La première composante principale peut être construite de façon à ce qu'elle soit corrélée positivement à l'ensemble des variables manifestes du bloc (du moins à une majorité d'entre elles). On rencontre un problème lorsque les VM sont négativement corrélées à la première composante principale.

b) Le α de Cronbach

Le α de Cronbach peut être utilisé afin de vérifier l'unidimensionnalité d'un bloc de p variables x_h lorsqu'elles sont positivement corrélées. Pour des variables standardisées, on l'obtient grâce à :

$$\alpha = \frac{\sum_{h \neq h'} cor(x_h, x_{h'})}{p + \sum_{h \neq h'} cor(x_h, x_{h'})} \times \frac{p}{p-1} \tag{3}$$

Le α de Cronbach peut aussi être défini pour des variables dans leur échelle originale par :

$$\alpha = \frac{\sum_{h \neq h'} cov(x_h, x_{h'})}{var\left(\sum_h x_h\right)} \times \frac{p}{p-1} \tag{4}$$

On considère généralement qu'un bloc est unidimensionnel lorsque le α de Cronbach est plus grand que 0,7.

c) Le ρ de Dillon-Goldstein

Le signe de la corrélation entre chaque VM et leur VL est connu par construction et on suppose qu'il est positif. Dans l'équation (1), cette hypothèse signifie que tous les loadings π_h sont positifs. Un bloc est unidimensionnel si tous ces loadings sont grands.

Le ρ de Dillon-Goldstein est défini par :

$$\rho = \frac{\left(\sum_{h=1}^{p} \pi_h\right)^2 \mathrm{Var}(\xi)}{\left(\sum_{h=1}^{p} \pi_h\right)^2 \mathrm{Var}(\xi) + \sum_{h=1}^{p} Var\left(\epsilon_h\right)} \tag{5}$$

Supposons que toutes les VM x_h et la VL ξ sont standardisées. Une approximation de la variable latente ξ peut être obtenue en standardisant la première composante principale t_1

associée à l'analyse en composantes principales sur le bloc de VM. Alors, π_h est estimé par $cor(x_h, t_1)$ et en utilisant l'équation (1), $Var(\epsilon_h)$ est estimé par $1 - cor^2(x_h, t_1)$. On obtient donc une estimation du ρ de Dillon-Goldstein :

$$\hat{\rho} = \frac{\left[\sum_{h=1}^{p} \text{cor}\left(x_h, t_1\right)\right]^2}{\left[\sum_{h=1}^{p} \text{cor}\left(x_h, t_1\right)\right]^2 + \sum_{h=1}^{p}\left[1 - \text{cor}^2\left(x_h, t_1\right)\right]} \tag{6}$$

Un bloc est supposé unidimensionnel lorsque le ρ de Dillon-Goldstein est plus grand que 0,7. Cette statistique constitue un meilleur indicateur que le α de Cronbach afin de juger de l'unidimensionnalité d'un bloc de VM (Chin, 1998, p.320).

L'approche LPS est un mélange de connaissance a priori et d'analyse de données. Lorsqu'on utilise la manière réflective, la connaissance a priori concerne l'unidimensionnalité des blocs et le signe des loadings. Les données doivent s'ajuster au modèle. Si celles-ci ne s'ajustent pas, il faudra retirer la variable manifeste qui pose problème. Une autre solution réside dans l'utilisation de la manière formative que nous allons décrire par la suite.

2.2. La manière formative

Dans le cas formatif, on suppose que la variable latente ξ est construite à partir de ses propres variables manifestes. La VL est une combinaison linéaire des variables manifestes associées en ajoutant un terme d'erreur :

$$\xi = \sum_h \varpi_h x_h + \delta \tag{7}$$

Dans le cas formatif, les blocs de variables peuvent être multidimensionnels. La seule hypothèse imposée est la suivante :

$$E\left(\xi | x_1, \ldots, x_{p_j}\right) = \sum_h \varpi_h x_h \tag{8}$$

Cette hypothèse implique que le vecteur de résidus δ a une moyenne de 0 et n'est pas corrélé aux VM x_h.

2.3. La manière MIMIC

La manière MIMIC est un mélange des manières formatives et réflectives.

La modèle de mesure pour un bloc est le suivant :

$$x_h = \pi_{h_0} + \pi_h \xi + \epsilon_h \text{ pour } h = 1, \ldots, p_1 \tag{9}$$

et

$$\xi = \sum_{h=p_1+1}^{p} \varpi_h x_h + \delta \tag{10}$$

Les p_1 premières variables sont réflectives et les $(p - p_1)$ dernières sont formatives. L'hypothèse de base utilisée reste la même que plus haut.

3. Le modèle structurel

Cette partie du modèle relie les variables latentes en utilisant des équations linéaires :

$$\xi_j = \beta_{j_0} + \sum_i \beta_{ji}\xi_i + v_j \tag{11}$$

Une variable latente qui n'est expliquée par aucune autre est appelée exogène. Dans le cas contraire, on l'appelle endogène.

4. L'algorithme d'estimation

4.1. Calcul des scores des variables latentes

Les variables latentes sont estimées en utilisant un algorithme itératif.

4.1.1. Estimation externe y_j des variables latentes standardisées ($\xi_j - m_j$)

Les variables latentes standardisées (moyenne = 0 et écart-type = 1) sont obtenues par combinaison linéaire des variables manifestes centrées :

$$y_j \propto \pm \left[\sum w_{jh}\left(x_{jh} - \overline{x}_{jh}\right)\right] \tag{12}$$

où le symbole "$\propto$" indique que le membre de gauche est égal au membre de droite standardisé et le symbole "$\pm$" montre qu'il existe une ambigüité sur le signe. On choisit le signe de façon à ce que y_j soit positivement corrélé avec le plus de VM x_{jh} possible.

La variable latente standardisée peut s'écrire :

$$y_j = \sum \tilde{w}_{jh}(x_{jh} - \overline{x}_{jh}) \tag{13}$$

Les coefficients w_{jh} et $\tilde{w}_{jh}$ sont appelés des poids externes.

La moyenne m_j est estimé par :

$$\hat{m}_j = \sum \tilde{w}_{jh}\overline{x}_{jh} \tag{14}$$

Et la variable latente ξ_j par :

$$\hat{\xi}_j = \sum \tilde{w}_{jh} x_{jh} = y_j + \hat{m}_j \tag{15}$$

Lorsque toutes les variables manifestes ont la même échelle de mesure, il est pratique d'exprimer les scores des variables latentes dans leur échelle d'origine (Fornell (1992)) :

$$\hat{\xi}_j^{\star} = \frac{\sum \tilde{w}_{jh} x_{jh}}{\sum \tilde{w}_{jh}} \tag{16}$$

L'équation (16) peut être calculée lorsque tous les poids externes sont positifs. Généralement, on utilise une échelle de 0 à 100 afin de comparer les scores des variables latentes, on écrira :

$$\hat{\xi}_j^{0-100} = 100 \times \frac{\hat{\xi}_j^{\star} - x_{min}}{x_{max} - x_{min}} \tag{17}$$

où x_{min} et x_{max} sont respectivement le minimum et le maximum de l'échelle de mesure commune à toutes les variables manifestes.

4.1.2. Estimation interne z_j des variables latentes standardisées ($x_j - m_j$)

L'estimation interne z_j des variables latentes standardisées ($x_j - m_j$) est définie par :

$$z_j \propto \sum_{j' : \xi_{j'} \text{ est reliée à } \xi_j} e_{jj'} y_{j'} \tag{18}$$

où les poids internes doivent être définis en choisissant un schéma de calcul.

Le schéma centroïde :

Le poids interne $e_{jj'}$ est égal au signe de la corrélation entre l'estimation externe y_j de la variable latente et celle $y_{j'}$ à condition que les variables latentes x_j et $x_{j'}$ soient reliées.

Ce schéma est le plus fréquemment utilisé, il a malheureusement un inconvénient : lorsque les corrélations sont très proches de 0, le signe peut changer lors de petites fluctuations. Néanmoins, dans des cas pratiques, ceci pose rarement problème.

Dans l'algorithme original, l'estimation interne n'est pas standardisée. Nous préférons la standardiser car ceci n'implique pas de changements et permet de simplifier certaines équations.

Le schéma factoriel :

Le poids interne e_{ji} est égal à la corrélation entre y_i et y_j. Ce schéma a été créé en réponse à l'inconvénient du schéma centroïde.

Le schéma structurel :

Les variables latentes connectées à ξ_j sont divisées en deux groupes : celles qui expliquent ξ_j et celles qui sont expliquées par ξ_j.

Pour une variable qui explique ξ_j, le poids interne est égal au coefficient de régression de y_j dans la régression multiple de y_j sur l'ensemble des estimations externes des prédécesseurs de ξ_j. Pour une variable qui est expliquée par ξ_j, le poids interne est égal à la corrélation entre les deux estimations externes associées aux deux variables latentes.

Ces nouveaux schémas n'ont pas une forte influence sur les résultats mais ils constituent des points théoriques importants. En effet, ils permettent de relier l'approche PLS à de nombreuses méthodes d'analyse de tableaux multiples.

Le schéma de Horst :

Le poids interne e_{ji} est toujours égal à 1. C'est l'un des premiers schémas développé pour l'approche PLS.

4.2. L'algorithme PLS d'estimation des poids

4.2.1. Les modes d'estimation des poids externes w_{jh}

Il y a trois manières classiques d'estimer les poids externes : le mode A, le mode B et le mode C.

Mode A :

Dans le mode A, les poids externes w_{jh} sont les coefficients de régression de z_j lorsqu'on fait une régression simple de x_{jh} sur l'estimation interne z_j de la variable latente ξ_j :

$$w_{jh} = cov(x_{jh}, z_{jh}) \tag{19}$$

car z_j est standardisée

Mode B :

Dans le mode B, le vecteur w_j des poids externes w_{jh} est le vecteur des coefficients de régression associé à la régression multiple de z_j sur les variables manifestes centrées ($x_{jh} - \overline{x}_{jh}$) associées à la même variable latente ξ_j :

$$w_{jh} = \left({X_j}' - X_j\right)^{-1} {X_j}' z_j \tag{20}$$

où X_j est une matrice dont les colonnes sont définies par les variables manifestes centrées $x_{jh} - \overline{x}_{jh}$ associées à la variable latente ξ_j.

Le mode A est adapté pour un bloc avec un modèle de mesure réflectif et mode B pour le cas formatif. Le mode A est fréquemment utilisé pour des variables latentes endogènes et le mode B lorsque celles-ci sont exogènes. Ces deux modes peuvent être utilisés simultanément dans le cas MIMIC.

Dans beaucoup de cas pratiques, le mode B est difficile à utiliser. En effet, il peut y a voir de fortes colinéarités à l'intérieur d'un bloc. Dans ce cas, on peut utiliser des régressions PLS à la place des régressions multiples OLS. On peut noter que le mode A revient à prendre la première composante d'une régression PLS et le mode B revient à prendre toutes les composantes de la régression PLS (on arrive ainsi à la régression linéaire multiple OLS).

Mode C (centroïde) :

Dans le mode centroïde, les poids externes sont tous égaux en valeur absolue et sont pris comme le signe de la corrélation entre les variables manifestes et leur variable latente :

$$w_{jh} = sign(cor(x_{jh}, z_j)) \tag{21}$$

Ces poids externes sont ensuite normalisés de façon à ce que la variable latente obtenue ait une variance de 1. Ce mode est associé à un modèle formatif et consiste en un cas particulier du mode B.

D'autres modes sont disponibles dans XLSTAT, le mode PLS avec utilisation de la régression PLS, le mode ACP avec l'utilisation de la première composante principale de l'analyse en composante principale appliquée sur le bloc et le mode MIMIC qui combine les modes A et B.

4.2.2. Estimation des poids

La première étape de l'algorithme PLS consiste en le choix arbitraire d'un vecteur de poids externes initiaux. Ces poids sont standardisés de façon à obtenir une variable latente de variance égale à 1.

Un bon choix pour les poids initiaux est de prendre $w_{jh} = sign(cor(x_{jh}, \xi_h))$ ou, plus simplement, $w_{jh} = sign(cor(x_{jh}, \xi_h))$ pour $h = 1$ et 0 sinon ou encore de prendre les éléments du premier vecteur propre de l'ACP sur chaque bloc.

Ensuite, les étapes d'estimations externes et internes sont répétées avec les modes et schémas prédéfinis jusqu'à convergence (celle-ci est prouvée uniquement pour le cas de deux blocs, au-delà elle n'est que constatée).

Après la dernière étape, le résultat final est atteint pour un poids externe noté $\tilde{w}_{jh}$. On calcule alors la variable latente standardisée $y_j = \sum \tilde{w}_{jh}(x_{jh} - \overline{x}_{jh})$, l'estimation de la moyenne $\hat{m}_j = \sum \tilde{w}_{jh}\overline{x}_{jh}$de la variable latente ξ_j et l'estimation finale du score $\hat{\xi}_j = \sum \tilde{w}_{jh}x_{jh} = y_j + \hat{m}_j$ de ξ_j. Cette dernière estimation peut être remise à l'échelle d'origine en utilisant les équations (16) et (17).

L'estimation des variables latentes sont sensibles à l'échelle des variables manifestes dans le cas du mode A, mais pas dans celui du mode B. Dans ce second cas, les estimations externes

des variables latentes sont les projections des estimations internes dans l'espace généré par ses variables manifestes.

4.3. L'estimation des équations structurelles

Les équations structurelles sont estimées en utilisant des régressions linéaires multiples classiques dans lesquelles les variables latentes sont remplacées par leurs scores estimés par l'algorithme PLS. Ce type de régression rencontre des problèmes lorsqu'il existe une certaine colinéarité entre les scores des variables latentes. Dans ce cas, on peut remplacer les régressions classiques par des régressions PLS.

5. Le traitement des données manquantes

XLSTAT-PLSPM permet de traiter les données manquantes d'une manière spécifique (Lohmöller, 1989):

1. Lorsque certaines données sont manquantes, les moyennes et écart-types des variables manifestes sont calculés sur les données disponibles.

2. Toutes les variables manifestes sont centres.

3. Si une observation est manquante pour toutes les variables associées à un bloc, alors la valeur de l'estimation externe est manquante pour cette observation.

4. Si une observation a quelques données manquantes associées à certaines variables d'un bloc j, alors l'estimation externe de la variable latente est définie par :

$$y_{ji} = \sum_{jh:x_{jhi} \text{ existe}} \tilde{w}_{jh}(x_{jhi} - \overline{x}_{jh})$$

Ceci revient à remplacer les données manquantes sur la variable x_{jh} par la moyenne $\overline{x}_{jh}$.

5. Si une observation a des données manquantes au niveau de la variable latente (c'est-à-dire qui'il n'y a pas de données pour l'ensemble des variables associées au bloc), alors l'estimation interne z_{ji} de la variable latente est définie par :

$$z_{ji} = \sum_{k:\xi_k \text{ est connecté avec } \xi_j \text{ et } y_{ki} \text{ existe}} e_{jk} y_{ki}$$

Ceci revient à remplacer les valeurs manquantes de y_k par la moyenne (c'est-à-dire 0).

6. Les poids externes sont calculés en utilisant l'ensemble des données disponibles avec la procédure suivante :

- Pour le mode A : le poids externe w_{jh} est le coefficient de corrélation de z_j dans la régression de $(x_{jh} - \overline{x}_{jh})$ sur z_j calculé sur les données disponibles.

- Pour le mode B : lorsqu'il n'y a pas de données manquantes, on peut écrire w_j :

$$w_j = [Var(X_j)]^{-1} Cov(X_j, z_j)$$

où $Var(X_j)$ est la matrice de covariance de X_j et $Cov(X_j, z_j)$ est le vecteur-colonne comprenant les covariances entre x_{jh} et z_j.

Lorsqu'il y a des données manquantes, chaque élément de $Var(X_j)$ et de $Cov(X_j, z_j)$ est calculé en utilisant toutes les paires de données disponibles et w_j est calculé en utilisant la formule précédente.

Le système par paires a l'inconvénient de permettre de calculer des covariances sur des échantillons de tailles différentes. Cependant, lorsqu'il y a peu de valeurs manquantes, cette méthode est très robuste. Ceci explique le fait que la procédure de « blindfolding » que nous présentons plus loin obtient des écart- types très petits pour les paramètres.

7. Les coefficients structurels sont les coefficients de régression issus de la régression multiple de certaines variables latentes sur d'autres. Lorsqu'il y a des données manquantes, la procédure introduite au point 6 est utilisée pour estimer les coefficients structurels.

D'autres méthodes plus classiques existent afin de traiter les données manquantes tel que l'imputation par la moyenne, la délétion par liste, l'imputation multiple ou encore l'algorithme NIPALS (dont nous parlerons plus loin).

6. Validation du modèle

Un modèle d'équations structurelles peut être validé à trois niveaux : (1) la qualité du modèle de mesure, (2) la qualité du modèle structurel, et (3) la qualité de chaque équation structurelle.

6.1. Communalité et redondance

L'indice de communalité mesure la qualité du modèle de mesure pour chaque bloc. Il est défini, pour le bloc j, par :

$$Communalité_j = \frac{1}{p_j} \sum_{h=1}^{p_j} cor^2(x_{jh}, y_j) \qquad (22)$$

La communalité moyenne est la moyenne de l'ensemble des $\text{cor}^2(x_{jh}, y_j)$:

$$\overline{Communalité} = \frac{1}{p} \sum_{j=1}^{J} p_j Communalité_j \qquad (23)$$

Où p est le nombre total de variables manifestes dans tous les blocs.

La redondance mesure la qualité du modèle structurel pour chaque variable latente endogène. Elle est définie, pour le bloc j, par :

$$Redondance_j = Communalité_j \times R^2(y_j, \{ \text{les } y_{j'} \text{ expliquent } y_j\}) \qquad (24)$$

La moyenne des redondances sur l'ensemble des variables latentes endogènes peut aussi être calculée.

Un critère global de qualité d'ajustement (GoF) existe (Amato, Esposito Vinzi and Tenenhaus, 2004). Il représente la moyenne géométrique de la communalité moyenne et du R² moyen :

$$GoF = \sqrt{\overline{\text{Communalite}} \times \overline{R^2}} \quad (25)$$

A la différence de la méthode par analyse de la structure de covariance (LISREL), l'approche PLS n'optimise aucun critère global, on ne peut donc pas obtenir d'indice permettant une validation global du modèle (à la différence du de la méthode LISREL). Le GoF représente une solution pratique utile car il juge simultanément de la qualité des modèles de mesure et structurel.

6.2. L'approche "blindfolding" : communalité et redondance par validation croisée

La cv-communalité (cv pour validation croisée en anglais) mesure la qualité du modèle de mesure pour chaque bloc. C'est une sorte de R^2 entre les VM du bloc et leur variable latente obtenu par validation croisée.

La qualité de chaque équation structurelle est mesurée grâce à la cv- redondance (le Q^2 de Stone et Geisser). C'est une sorte de R^2 obtenu par validation croisée entre les variables manifestes associées à une variable latente endogène et les variables manifestes associées aux variables latentes expliquant la variable latente endogène en se basant sur le modèle structurel estimé.

Le niveau de significativité des coefficients de régression peut être calculé en utilisant la statistique t de Student habituelle ou des méthodes de validation croisées telles que le bootstrap et le jack-knife.

Nous présentons une description de la méthode de « blindfolding » développée par Herman Wold :

1. La matrice des données est divisée en G groupes. Une valeur de 7 est recommandée par Herman Wold. Nous rassemblons dans le tableau suivant un exemple de jeu de données avec 12 observations et 5 variables. Le premier groupe est associé à la lettre a, le second à b et ainsi de suite.

x_1 | x_2 | x_3 | x_4 | x_5 |----|----|----|----|----| a|f|d|b|g| b|g|e|c|a| c|a|f|d|b| d|b|g|e|c| e|c|a|f|d| f|d|b|g|e| g|e|c|a|f| a|f|d|b|g| b|g|e|c|a| c|a|f|d|b| d|b|g|e|c| e|c|a|f|d|

2. Chacun à son tour, chaque groupe de cellules est retiré du jeu de données.

3. Le modèle PLS associé est estimé. On répète alors cette estimation G fois.

4. L'un des moyens pour évaluer la qualité du modèle réside dans le fait de mesurer la capacité à prédire une variable manifeste en utilisant les variables latentes. Deux indices sont utilisés : la communalité et la redondance.

5. Pour la communalité, on obtient une prédiction des variables manifestes centrées qui n'ont pas été inclues dans l'analyse en utilisant les estimations des variables latentes et la formule suivante :

$$Pred(x_{jhi} - \overline{x}_{jh}) = \hat{\pi}_{jh(-i)} y_{j(-i)}$$

où $\hat{\pi}_{jh(-i)}$ et $y_{j(-i)}$ sont calculés sur les données lorsque la i-ème valeur de la variable x_{jh} est manquante.

Les termes suivant sont calculés :

- La somme des carrés des observations pour une VM : $SSO_{jh} = \sum_i (x_{jhi} - \overline{x}_{jh})^2$.
- La somme des carrés des erreurs de prévision pour une VM : $SSE_{jh} = \sum_i (x_{jhi} - \overline{x}_{jh} - \hat{\pi}_{jh(-i)} y_{j(-i)})^2$.
- La somme des carrés des observations pour le bloc j : $SSO_j = \sum_h SSO_{jh}$.
- La somme des carrés des erreurs de prévision pour le bloc j : $SSE_j = \sum_h SSE_{jh}$.
- La cv-communalité pour le bloc j : $H_j^2 = 1 - \frac{SSE_j}{SSO_i}$.

Le H_j^2 est la communalité obtenue par validation croisée.

6. Pour la redondance, on obtient une prédiction des variables manifestes centrées qui n'ont pas été inclues dans l'analyse en utilisant :

$$Pred(x_{jhi} - \overline{x}_{jh}) = \hat{\pi}_{jh(-i)} Pred(y_{j(-i)})$$

où $\hat{\pi}_{jh(-i)}$ est défini de la même façon que dans le paragraphe précédent et $Pred(y_{j(-i)})$ est la prédiction pour la i-ème observation de la variable latente en utilisant le modèle de régression lorsque la i-ème valeur de la variable est manquante.

On calcule alors :

- La somme des carrés des erreurs de prévision pour une VM :

$$SSE'_{jh} = \sum_i (x_{jhi} - \overline{x}_{jh} - \hat{\pi}_{jh(-i)} Pred(y_{j(-i)}))^2$$

- La somme des carrés des erreurs de prévision pour le bloc j :

$$SSE'_j = \sum_h SSE'_{jh}$$

- La cv-redondance pour le bloc endogène j :

$$F_j^2 = 1 - \frac{SSE'_j}{SSO_j}$$

Le F_j^2 est la redondance obtenue par validation croisée.

6.3. Rééchantillonnage : Jackknife et Bootstrap

La significativité des paramètres du modèle peut être obtenue en utilisant des procédures non paramétriques. Mise à part la validation croisée, on peut aussi utiliser le bootstrap et le jackknife.

6.3.1. Jackknife

La procédure de rééchantillonnage est basée sur la suppression d'un certain nombre d'observation (en général 1) de l'échantillon original (de taille N). Chaque sous-échantillon a alors $N-1$ observations. Si on augmente le nombre d'observations supprimées, une perte de la robustesse de la statistique t est possible. La procédure complète est décrite dans Chin (1998, p.318-320).

6.3.2. Bootstrap

Le rééchantillonnage bootstrap est basé sur des triages avec remise des observations de l'échantillon original. Les échantillons obtenus ont N observations. Il faut spécifier le nombre de répétitions (fixé par défaut à 100).

Il faut prendre en compte le fait que dans l'approche PLS, le signe des variables latentes n'est pas défini. Ceci revient à dire que $y_j = \sum \tilde{w}_{jh}(x_{jh} - \overline{x}_{jh})$ et $-y_j$ sont des solutions équivalentes. Il existe deux méthodes principales afin d'éviter ce problème: Wold (1985) propose de garder la solution pour laquelle y_j est corrélé positivement au plus de variables manifestes x_{jh} possible. La seconde approche consiste à retenir le signe de la première valeur propre obtenue sur l'échantillon original.

GSCA (Generalized Structured Component Analysis)

Cette méthode introduite par Hwang et Takane (2005) permet d'optimiser une fonction globale en utilisant un algorithme de moindres carrés alternés (ALS).

GSCA se trouve dans la tradition de l'analyse en composantes. Il remplace les facteurs par des composantes comme dans l'approche PLS. Néanmoins, GSCA propose un critère d'optimisation global, qui est minimisé afin d'obtenir les paramètres du modèle. Cette méthode a donc un indice global de qualité d'ajustement tout en gardant tous les avantages de l'approche PLS.

Soit Z une matrice N par J des variables observées. On suppose que Z est centrée et réduite. Donc le model GSCA peut s'écrire :

$$ZV = ZWA + E$$

et

$$P = GA + E \tag{1}$$

avec $P = ZV$, et $G = ZW$. Dans (1), P est une matrice N par T des variables observées endogènes et des variables composites, G est une matrice N par D des variables exogènes observées et des variables composites. V est une matrice J par T des poids associés aux variables endogènes, W est une matrice J D des poids des variables exogènes, A est une super-matrice D par T composée des matrices des loadings reliant les composantes aux variables observées, notée C, associée à la matrice des coefficients structurels, notée B, on a, $A = [C, B]$, et E est une matrice de résidus.

On estime les inconnues V, W, et A de manière à ce que la somme des carrés des résidus, $E = ZV - ZWA = P - GA$, soit aussi petite que possible. Ceci revient à minimiser :

$$f = SS(ZV - ZWA) = SS(P - GA) \tag{2}$$

Par rapport à V, W, and A, où $SS(X) = trace(X'X)$. Les composantes dans P et/ou G doivent être normalisées pour des problèmes d'identification.

On ne peut pas résoudre de manière analytique cette équation. On utilise donc un algorithme de moindres carrés alternés (ALS) (de Leeuw, Young, & Takane, 1976) afin de minimiser (2). En général, l'algorithme ALS peut être vu comme un cas spécial d'algorithme du point fixe (FP) dans lequel le point fixe est un point stationnaire d'une fonction à optimiser.

L'algorithme proposé se sépare en deux étapes : Dans la première étape, A est mis à jour pour V et W fixés. Dans une seconde étape, V et W sont mis à jour pour un A fixé. (Hwang and Takane, 2004)

RGCCA (Regularized Generalized Canonical Correlation Analysis)

Cette méthode issue de Tenenhaus et al. (2011) permet de se rapprocher de l'approche PLS en gardant des fonctions à optimiser (à la différence de l'approche PLS).

A la différence de l'approche PLS, les résultats de la méthode RGCCA sont des corrélations entre les variables latentes et entre les variables manifestes et leurs variables latentes associées (il n'y a pas de régressions à la fin de l'algorithme).

La méthode RGCCA est basée sur un algorithme itératif simple proche de celui de l'approche PLS qui se décompose de la manière suivante :

1 - Initialisation des poids externes de la même façon que dans l'algorithme PLSPM.

2 - Normalisation des poids externes en utilisant le paramètre tau :

$$w_j^0 = \left[\left(w_j^0\right)^T \left[\tau_j I + (1-\tau_j) \frac{1}{n} X_j^T X_j \right]^{-1} w_j^0 \right]^{-1/2} \left[\tau_j I + (1-\tau_j) \frac{1}{n} X_j^T X_j \right]^{-1} w_j^0$$

3 - Calcul des composantes internes de chaque variable latente en fonction du schéma utilisé (les schémas sont les mêmes qu'en PLSPM)

$$z_j^s = \sum_{k<j} c_{jk} e_{jk} X_k w_k^{s+1} + \sum_{k>j} c_{jk} e_{jk} X_k w_k^s$$

Avec e_{jk} poids interne et $c_{jk} = 1$ si les variables latentes j et k sont liées.

4 - Mise à jour des poids externes :

$$w_j^{s+1} = \left[\left(z_j^s\right)^T X_j \left[\tau_j I + (1-\tau_j)\frac{1}{n} X_j^T X_j\right]^{-1} X_j^T w_j^s \right]^{-1/2} \left[\tau_j I + (1-\tau_j)\frac{1}{n} X_j^T X_j\right]^{-1} X_j^T w_j^s$$

5- On répète les étapes 3 et 4 jusqu'à convergence de l'algorithme.

Une fois que l'algorithme a convergé, on obtient des résultats qui optimisent des fonctions bien précises en fonction du choix du paramètre tau.

Ce paramètre permet d'ajuster le « mode ». Si tau=0 alors on sera dans le cas du mode B et les résultats de la méthode RGCCA et de l'approche PLS concordent complètement. Lorsque tau=1, on se trouve dans le cas de ce qu'appelle M. Tenenhaus, le nouveau mode A qui est proche du mode A tout en optimisant une fonction donnée. Lorsque tau varie entre 0 et 1, on se rapproche plus ou moins de l'un des deux modes. Pour plus de détails sur la méthode RGCCA, voir Tenenhaus et al. (2011).

Dans le cadre de la méthode RGCCA, XLSTAT-PLSPM propose aussi le mode Ridge RGCCA. Dans ce cas, le paramètre tau est optimisé de manière à définir le mieux possible chaque bloc. Pour calculer la valeur du paramètre, la formule de Schäfer et Strimmer (2005) reproduite dans Tenenhaus et al. (2011) est utilisée.

L'algorithme NIPALS

L'algorithme PLS est issu de l'algorithme NILES (Non linear Iterative LEast Squares estimation), qui est devenu plus tard l'algorithme NIPALS (Non linear Iterative PArtial Least Squares) afin d'effectuer une analyse en composantes principales avec des données manquantes (Wold, 1966). Cet algorithme a deux intérêts dans le cadre de l'approche PLS : la prise en compte de données manquantes et la possibilité de travailler sur plusieurs dimensions.

L'algorithme original, afin d'être intégré dans l'approche PLS, doit être légèrement modifié : on standardise les composantes principales. Une fois cette étape effectuée, la dernière étape de l'algorithme NIPALS revient à l'application du mode A de l'approche PLS avec un seul bloc de variables. Ceci revient à dire que l'approche PLS permet d'obtenir les résultats du premier ordre de l'analyse en composantes principales lorsque la manière réflective est sélectionnée.

Les dimensions supplémentaires sont obtenues en traitant les résidus de X avec les composantes principales standardisées obtenues précédemment.

L'approche PLS dans le cas de deux blocs de variables

L'approche PLS peut être mise ne relation avec de nombreuses méthodes d'analyses de données lorsqu'il existe deux blocs de variables. Le tableau suivant rassemble les équivalences

entre l'approche PLS et d'autres méthodes d'analyse de données lorsque deux blocs sont traités. Il y est précisé l'utilisation de méthodes de déflation pour le cas de plusieurs dimensions.

--	Analyse canonique des corrélations	Analyse factorielle inter-batteries	Régression PLS de X2 sur X1	Analyse de redondances de X2 par rapport à X1
Mode associé à X1	B (déflation)	A (déflation)	A (déflation)	B (déflation)
Mode associé à X2	B (déflation)	A (déflation)	A (pas de déflation)	A (pas de déflation)

On peut trouver les démonstrations de ces équivalences dans Tenenhaus et al. (2005).

L'approche PLS dans le cas de J groupes de variables

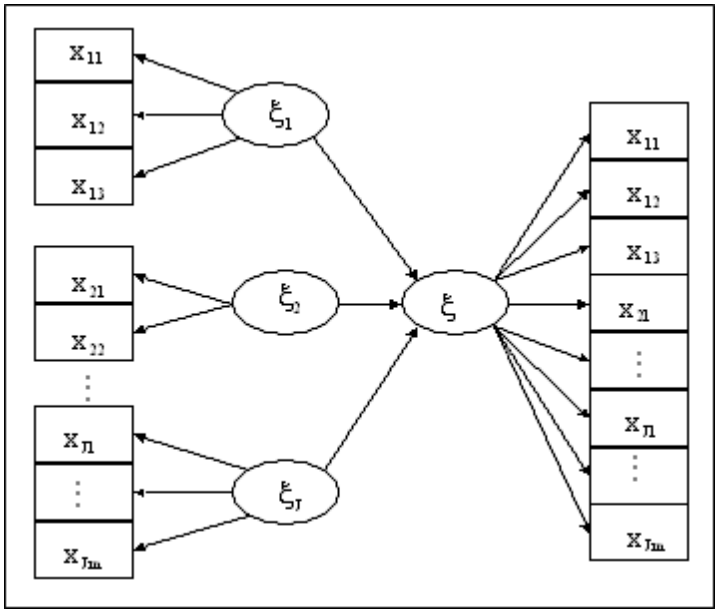
Les diverses options de l'approche PLS (modes A ou B pour l'estimation externe ; les schémas centroïde, factoriel ou structurel pour l'estimation interne) permettent de retrouver de nombreuses méthodes d'analyse multi-tableaux : l'analyse canonique généralisée (de Horst (1961) et de Carroll (1968)), l'analyse factorielle multiple (Escofier et Pagès, 1994), l'analyse en composantes principales « split » de Lohmöller (1989), l'algorithme de variance maximale de Horst (1965).

Le lien entre ces méthodes et l'approche PLS a été étudié dans le cas d'exemples pratiques dans Guinot, Latreille et Tenenhaus (2001) et dans Pagès et Tenenhaus (2001).

Soit J blocs de variables observées $X_1, \ldots, X_j$ sur les mêmes observations. Afin d'estimer les variables latentes ξ_j, Wold (1982) a proposé le modèle hiérarchique suivant :

- Un nouveau bloc X est construit en rassemblant les J blocs $X_1, \ldots, X_j$ dans un super bloc
- Le super bloc X est associé à une seule variable latente ξ.
- L'ensemble des variables latentes exogènes ξ_j sont reliées à la variable latente endogène ξ.

La figure suivante illustre le cas de 3 blocs :



.

Le tableau suivant rassemble les correspondances entre l'approche PLS hiérarchique et mes méthodes de tableaux multiples en fonction des schémas d'estimation.

Mode pour l'estimation externe	Schéma centroïde	Schéma factoriel	Schéma structurel
A	Analyse canonique de corrélations généralisée PLS de Horst	Analyse canonique de corrélations généralisée PLS de Carroll	Analyse en composante principale « split » de Lohmöller / Algorithme de la variance maximale de Horst / Analyse factorielle multiple d'Escoffier et Pagès
B	Analyse canonique de corrélations généralisée de Horst (critère SUMCOR)	Analyse canonique de corrélations généralisée de Carroll	

Dans le cadre des méthodes décrites dans le tableau précédent, les dimensions supplémentaires peuvent être obtenues en appliquant à nouveau l'approche PLS après une déflation du bloc X.

On peut aussi obtenir des composantes orthogonales de plus grandes dimensions pour certains blocs X_j. L'approche PLS hiérarchique est appliquée à nouveau après déflation des blocs concernés.

Le contrôle de l'orthogonalité constitue un avantage important de l'approche PLS (on peut voir Tenenhaus (2004) pour plus de détails et pour une application).

Finalement, l'approche PLS peut être vue comme une méthode générale pour l'analyse des tableaux multiples. Il a été démontré que l'on pouvait retrouver la majorité des méthodes

classiques d'analyse des tableaux multiples, mais il est possible de construire de nouvelles méthodes en faisant varier les différents paramètres associés aux différentes estimations du modèle. Ainsi, on peut dire que l'approche PLS est un outil très flexible afin d'étudier des tableaux multiples en utilisant des relations structurelles à variables latentes.

Les tests de comparaison multigroupes dans le cadre de l'approche PLS

Il peut être intéressant de comparer des groupes d'observations sur un modèle donné. Des tests permettent la comparaison des path coefficients ou d'autres indices. Nous utilisons deux types de tests : une adaptation du test t utilisant les écart-types bootstrap et un test de permutation.

Le test t multigroup :

Wynne Chin a été le premier à utiliser ce test afin de comparer des path coefficients. Ce test est basé sur la différence entre les path coefficients et suit une distribution de Student. On définit ainsi la statistique t :

$$t = \frac{\left|\beta_{ij}^{G_1} - \beta_{ij}^{G_2}\right|}{\sqrt{\frac{(n_1-1)^2}{n_1+n_2-2}SE_{G_1}^2 + \frac{(n_2-1)^2}{n_1+n_2-2}SE_{G_2}^2} \cdot \sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$$

où n_1 et n_2 sont les tailles respectives des deux groupes et où $SE_{G_i}^2$ est la variance du coefficient β_{ij} obtenue par bootstrap. Cette statistique suit une loi de Student à $n_1 + n_2 - 2$ degrés de liberté. Ce test fonctionne bien lorsque les données ne dévient pas trop de la normalité et que les variances des deux groupes sont proches.

Les tests de permutation :

Les tests de permutation offrent une bonne alternative adaptée au principe non paramétrique de l'approche PLS. Ils ont été utilisés dans le cadre de l'approche PLS par Chin (2008) et par Jakobowicz (2007). Leur principe est simple :

On sélectionne une statistique S. Dans le cas de l'approche PLS, on prendra la différence en valeur absolue entre deux paramètres.

On calcule la valeur de cette statistique sur les deux échantillons représentant les groupes jusqu'à S_{obs}.

On permute aléatoirement les éléments des deux groupes d'observations et on recalcule la statistique S_{permi}. On répète cette étape N_{perm} fois (avec N_{perm} très grand).

La p-valeur est telle que :

$$p - \text{valeur} = \frac{1}{N_{pem}+1}\sum_{i=1}^{N_{pen}} I\left(S_{obs} < S_{pem_i}\right)$$

La fonction $I(.)$ vaut 1 lorsque $S_{obs} < S_{permi}$ et 0 sinon.

La segmentation avec l'algorithme REBUS (Esposito Vinzi et al., 2008)

Il arrive qu'il existe une hétérogénéité au niveau des observations. Sur un modèle il existe différents groupes d'observations qui n'ont pas le même comportement. Dans ce cas, il peut être intéressant de rechercher s'il existe des groupes d'observations ayant un comportement semblable sur un modèle prédéfini. L'algorithme REBUS-PLS (*Response Based procedure for detecting unit segments in PLS path modelling*) permet de trouver des classes en utilisant le modèle PLS. Cette approche est basée sur un algorithme simple :

1- Application de l'approche PLS sur l'échantillon complet

2- Calcul des résidus associés aux variables manifestes et aux variables latentes

3- Application d'une classification ascendante hiérarchique (CAH) sur les résidus.

4- Application de l'approche PLS sur les modèles locaux

5- Calcul des résidus des variables manifestes et latentes et de l'indice de dissimilarité (CM index) pour chaque classe et chaque observation

6- Allocation des individus aux classes

7- Répéter 4 à 6 jusqu'à obtenir des classes stables.

L'indice CM index est donné par :

$$CM_{ik} = \sqrt{\frac{\sum_{j}\sum_{h=1}^{p_j}\left[e_{ihjk}^2 / \operatorname{Com}\left(\xi_{jk}, \mathbf{x}_{hj}\right)\right]}{\frac{\sum_{i=1}^{N}\sum_{j}\sum_{h=1}^{p_j}\left[e_{ihjk}^2 / \operatorname{Com}(\xi_{jk}, \mathbf{x}_{hj})\right]}{N - m_k - 1}} \times \frac{\sum_{j^*=1}^{J^*}\left[f_{ij^*k}^2 / R^2\left(\xi_{j^*}, \xi_{j:\xi_j \to \xi_{j^*}}\right)\right]}{\frac{\sum_{i=1}^{N}\sum_{j^*=1}^{J^*}\left[f_{ij^*k}^2 / R^2\left(\xi_{j^*}, \xi_{j:\xi_j \to \xi_{j^*}}\right)\right]}{N - m_k - 1}}}$$

Pour l'individu i et la classe k. On a e_{ihjk}, résidu obtenu à partir de la variable manifeste x_{hj} associé à la variable latente ξ_j, $Com()$ est la communalité, f_{ij^*k} le résidu obtenu à partir de la variable latente ξ_{j^*}. N est le nombre total d'observations et m_k est toujours égal à 1.

XLSTAT-PLSPM propose la méthode REBUS avec un certain nombre d'options :

- Le nombre de classes : Celui-ci peut être fixé manuellement ou calculé automatiquement dans le cadre de la Classification Ascendante Hiérarchique.

- La convergence : Un seuil en pourcentage doit être défini. Par exemple, si on prend 95 %, si 95 % des observations ne changent pas de classe d'une itération à la suivante alors on considère que le modèle est stable.

La méthode REBUS ne s'applique que si tous les blocs sont construits avec le mode A et si aucun groupe n'est sélectionné.

L'indice de qualité globale du modèle permet de juger de la qualité de la segmentation. Cet indice est équivalent au GoF lorsqu'une seule classe est utilisée. Lorsque plusieurs classes sont obtenues, il est obtenu de la manière suivante :

$$GQI = \sqrt{\sum_{k=1}^{K} \frac{n_k}{N} \left[\frac{1}{\sum P_q} \sum_{q} \sum_{p=1}^{P_q} \left(1 - \frac{\sum_{i=1}^{n_k} e_{ipqj}^2}{\sum_{i=1}^{n_k} \left(x_{ipq} - \overline{x}_{pqk} \right)^2} \right) \right]} \times \sqrt{\sum_{k=1}^{K} \frac{n_k}{N} \left[\frac{1}{J} \sum_{j=1}^{J} \left(1 - \frac{\sum_{i=1}^{n_k} f_{ijk}^2}{\sum_{i=1}^{n_k} \left(\hat{\xi}_{ipq} - \overline{\hat{\xi}}_{pqk} \right)^2} \right) \right]}$$

Avec n_k taille de la classe k, P_q nombre de variable manifestes associées à la variable latente q. e^2 résidus obtenus à partir des variables manifestes et f^2 résidus obtenus à partir des scores des variables latentes.

XLSTAT-PLSPM affiche en plus du GQI global, l'amélioration du GQI par rapport au modèle à une classe et la décomposition par rapport au modèle externe et au modèle interne du GQI.

L'affichage Market pour l'analyse de la satisfaction des consommateurs

Outre les modes d'affichage basique et avancé, XLSTAT-PLSPM propose un mode d'affichage appelé Market. Ce mode d'affichage propose une manière beaucoup plus simple de paramétrer le modèle et des sorties adaptées tout spécialement aux analyses dans le domaine du marketing et de l'analyse de la satisfaction.

Ce mode d'affichage fixe un certain nombre de paramètres par défaut et propose des traitements des variables simplifiés. On sélectionne tout d'abord l'échelle des variables manifestes et des variables latentes (voir partie Ajuster le modèle).

Par défaut, la méthode PLS, le schéma structurel et les régressions OLS sont utilisés. Les poids initiaux sont obtenus avec les valeurs du premier vecteur propre de l'analyse en composantes principales et une seule dimension est affichée.

L'option segmentation permet de faire de l'analyse REBUS pour extraire des groupes d'observations homogènes et les tests multigroupes peuvent être définis dans l'onglet options.

Il propose, de plus, des graphiques et des tableaux de simulation permettant d'observer l'impact sur une variable latente cible de modifications sur les variables manifestes et latentes du modèle.

Le chargement de modèle à partir de la librairie de modèles

XLSTAT-PLSPM vous permet de charger des modèles structurels existants de manière simple. Il vous suffit d'utiliser l'icône associé au chargement d'un modèle et de choisir de charger un modèle à partir de la bibliothèque de modèles.

Les modèles sont chargés à la place du modèle en cours tout en conservant vos données et vos résultats. Il vous faut ensuite associer les variables manifestes aux variables latentes.

Voici quelques-uns des modèles présents dans la bibliothèque sous forme d'un fichier .ppmxmod :

- ECSI (European Customer Satisfaction Index)
- ACSI (American Customer Satisfaction Index)
- SCSB (Swedish Customer Satisfaction Barometer)
- Norwegian Customer Satisfaction Barometer (NCSB),
- Swiss Index of Customer satisfaction (SWICS)
- Korean Customer Satisfaction Index (KCSI)
- Malaysian Customer Satisfaction Index(MCSI)
- ECSI simplifié (sans les réclamations)

Projets

Les projets XLSTAT-PLSPM sont des classeurs Excel particuliers. Lorsque vous créez un nouveau projet, son nom par défaut commence par PLSPMBook. Vous pouvez ensuite le sauvegarder sous un nom de votre choix, mais veillez à bien utiliser les boutons "Enregistrer" ou "Enregistrer sous" de la barre d'outils XLSTAT-PLSPM pour les enregistrer dans le répertoire dédié aux projets PLSPM, en utilisant l'extension *.ppm jusqu'à Excel 2003 et *.ppmx à partir de Excel 2007.

Un projet brut XLSTAT-PLSPM contient toujours deux feuilles qui ne doivent pas être supprimées :

- D1 : cette feuille est vide, et vos données doivent y être copiées/collées.

- PLSPMGraph : cette feuille est vide au départ, et doit être utilisée pour créer le modèle. Lorsque vous sélectionnez cette feuille, la barre d'outils "Path modeling" est affichée. Cette dernière est rendue invisible lorsque vous quittez cette feuille.

Une fois qu'un modèle a été créé, vous pouvez lancer l'estimation des paramètres du modèle. Les résultats sont ensuite affichés dans des feuilles Excel, à la suite de la feuille PLSPMGraph.

Il est possible d'enregistrer un modèle avant de le modifier, afin de pouvoir éventuellement le modifier par la suite (voir la section « Barres d'outils » pour plus de détails).

Lors de la création d'un nouveau projet, il vous sera demandé quel affichage vous désirez utiliser. Vous aurez le choix entre 3 options suivant vos objectifs et votre expertise.

Il est aussi possible de charger des modèles déjà enregistrés (voir la section « Barres d'outils » pour plus de détails).

Options

Pour afficher la boîte de dialogue des options, cliquez sur le bouton  de la barre d'outils "XLSTAT-PLSPM". Utilisez cette boîte de dialogue pour définir les options générales du module XLSTAT-PLSPM.

Onglet **Général** :

Mode d'affichage : L'approche PLS et le module XLSTAT-PLSPM étant complexes, trois types d'affichages sont possibles. L'affichage classique est celui utilisé par défaut, il permet d'effectuer les principales analyses de l'approche PLS. Le second permet d'afficher l'ensemble des fonctionnalités avancées. Finalement, l'affichage Market permet d'obtenir une paramétrisation adaptée aux analyses marketing et à l'analyse de la satisfaction.

Chemin pour les projets XLSTAT-PLSPM : ce chemin peut être modifié si et seulement si vous accès en lecture/écriture au chemin en question. Vous pouvez modifier le chemin en cliquant sur le bouton [...] puis en choisissant le dossier adéquat. Ce dossier doit être accessible en lecture/écriture.

Onglet **Format** :

Utilisez ces options pour définir le format des différents objets qui sont affichés sur la feuille PLSPMGraph :

- **Variables latentes** : vous pouvez choisir la couleur et l'épaisseur de la bordure des ellipses qui correspondent aux variables latentes, de même que la couleur du fond et, la couleur et la taille de la police.
- **Variables manifestes** : vous pouvez choisir la couleur et l'épaisseur de la bordure des rectangles qui correspondent aux variables manifestes, de même que la couleur du fond, et, la couleur et la taille de la police.
- **Flèches (MV-LV)** : vous pouvez choisir la couleur et l'épaisseur des flèches reliant les variables manifestes aux latentes.
- **Flèches (LV-LV)** : vous pouvez choisir la couleur et l'épaisseur des flèches reliant les variables latentes entre elles.

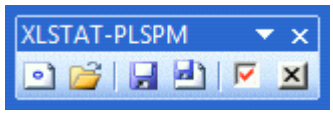
Remarque 1: pour que les changements soient effectifs vous devez cliquer sur le bouton OK, puis cliquer sur le bouton  de la barre « Path modeling ».

Remarque 2: ces options ne vous empêchent pas de changer le format d'un ou de plusieurs objets sur la feuille PLSPMGraph. En utilisant la barre de dessin d'Excel vous pouvez facilement modifier la couleur du fond ou des bordures des objets.

Barres d'outils

XLSTAT-PLSPM dispose de deux barres d'outils, "XLSTAT-PLSPM" et "Path modeling".

La barre d'outils "XLSTAT-PLSPM" peut être affichée en cliquant sur le bouton de la barre XLSTAT.



Cliquez sur ce bouton pour ouvrir un nouveau projet PLSPM (voir la section Projets pour plus de détails).

Cliquez sur ce bouton pour ouvrir un projet PLSPM existant.

Cliquez sur ce bouton pour enregistrer le projet PLSPM actif. Ce bouton n'est accessible que si des modifications ont été effectuées dans le projet.

Cliquez sur ce bouton pour enregistrer le projet dans un nouveau dossier ou sous un autre nom.

Cliquez sur ce bouton pour afficher la boîte de dialogue des options XLSTAT-PLSPM.

Cliquez sur ce bouton si vous souhaitez continuer à utiliser XLSTAT mais pas XLSTAT-PLSPM. Ferme XLSTAT-PLSPM permet de libérer de la mémoire.

EXCEL 2003 et antérieures

La seconde barre d'outils, "Path modeling" est uniquement visible lorsque vous êtes sur la feuille PLSPMGraph d'un projet PLSPM.

Cliquez sur ce bouton pour ajouter des variables latentes. Si vous double-cliquez sur ce bouton, vous pouvez ensuite ajouter plusieurs variables latentes à la suite, sans avoir à recliquer sur ce bouton.

Cliquez sur ce bouton pour ajouter des liens entre les variables latentes. Si vous double-cliquez sur ce bouton, vous pouvez ensuite ajouter plusieurs liens à la suite, sans avoir à recliquer sur ce bouton. Lorsque vous ajouter un lien, sélectionnez d'abord la variable latente qui sera à l'origine de la flèche, puis glissez le curseur de la souris jusqu'à la variable qui se trouvera à l'extrémité finale (la pointe) de la flèche.

Cliquez sur ce bouton pour afficher les variables manifestes. Si une variable latente est sélectionnée lorsque vous cliquez sur ce bouton, seules ses variables manifestes seront affichées.

Cliquez sur ce bouton pour ne plus afficher les variables manifestes. Si une variable latente est sélectionnée lorsque vous cliquez sur ce bouton, seules ses variables manifestes seront cachées.

Cliquez sur ce bouton pour optimiser l'affichage.

Cliquez sur ce bouton pour définir des groupes. Une fois que des groupes sont définis, une liste avec les libellés des groupes est affichée sur la feuille PLSPMGraph. Cette icône devient alors ; cliquez sur ce bouton pour ne plus tenir compte des groupes.

Cliquez sur ce bouton pour sauvegarder le modèle actuel dans le projet sous un nom de votre choix.

Cliquez sur ce bouton pour recharger un modèle préalablement sauvegardé.

Cliquez sur ce bouton pour supprimer un ou plusieurs modèles préalablement sauvegardés.

Déprotégé/Protégé(1)/Protégé(2): La première option permet à l'utilisateur de modifier le modèle et la position des objets. La seconde option permet de modifier uniquement la position des objets. La troisième option ne permet pas à l'utilisateur de modifier quoi que ce soit.

Cliquez sur ce bouton pour supprimer tous les objets de la feuille PLSPMGraph.

Cliquez sur ce bouton pour afficher les résultats de l'estimation des paramètres du modèle, si elle a déjà été effectuée. Si les résultats sont déjà affichés, le bouton suivant est affiché: ; cliquez alors ce bouton pour cacher les résultats.

Cliquez sur ce bouton pour afficher la boîte de dialogue des options d'affichage des résultats sur la feuille PLSPM.

Cliquez sur ce bouton pour démarrer l'optimisation du modèle puis pour afficher les résultats dans les feuilles de résultats et sur la feuille PLSPMGraph.

EXCEL 2007 et ultérieures

La seconde barre d'outils, "Path modeling" est uniquement visible lorsque vous êtes sur la feuille PLSPMGraph d'un projet PLSPM.

Cliquez sur ce bouton pour ajouter des variables latentes. Si vous double-cliquez sur ce bouton, vous pouvez ensuite ajouter plusieurs variables latentes à la suite, sans avoir à recliquer sur ce bouton.

Cliquez sur ce bouton pour ajouter des variables manifestes à la variable latente sélectionnée, ou utilisez le raccourci clavier Ctrl+M.

 Cliquez sur ce bouton pour afficher les variables manifestes. Si une variable latente est sélectionnée lorsque vous cliquez sur ce bouton, seules ses variables manifestes seront affichées.

 Cliquez sur ce bouton pour ne plus afficher les variables manifestes. Si une variable latente est sélectionnée lorsque vous cliquez sur ce bouton, seules ses variables manifestes seront cachées.

 Cliquez sur ce bouton pour optimiser l'affichage.

 Cliquez sur ce bouton pour définir des groupes. Une fois que des groupes sont définis, une liste avec les libellés des groupes est affichée sur la feuille PLSPMGraph. Cette icône devient alors  ; cliquez sur ce bouton pour ne plus tenir compte des groupes.

 Cliquez sur ce bouton pour sauvegarder le modèle actuel dans le projet sous un nom de votre choix.

 Cliquez sur ce bouton pour recharger un modèle préalablement sauvegardé ou charger un modèle issu d'une librairie de modèles (voir partie description).

 Cliquez sur ce bouton pour supprimer un ou plusieurs modèles préalablement sauvegardés.

Déprotégé/Protégé(1)/Protégé(2): La première option permet à l'utilisateur de modifier le modèle et la position des objets. La seconde option permet de modifier uniquement la position des objets. La troisième option ne permet pas à l'utilisateur de modifier quoi que ce soit.

 Cliquez sur ce bouton pour ajouter des liens entre les variables latentes. Sélectionnez d'abord la variable latente qui sera à l'origine de la flèche, puis celle qui se trouvera à l'extrémité finale (la pointe) de la flèche (utilisez la touche Ctrl ou Shift) et cliquez sur ce bouton, ou utilisez le raccourci clavier Ctrl+L. Pour inverser le sens utilisez le raccourci clavier Ctrl+R

 Cliquez sur ce bouton pour créer des liens à double sens entre toutes les variables. De manière équivalente, vous pouvez utiliser le raccourci clavier Ctrl+D.

 Cliquez sur ce bouton pour modifier la position des variables manifestes. De manière équivalente, vous pouvez utiliser le raccourci clavier Ctrl+O.

 Cliquez sur ce bouton renommer une variable latente après l'avoir sélectionnée.

 Cliquez sur ce bouton pour supprimer tous les objets de la feuille PLSPMGraph.

 Cliquez sur ce bouton pour afficher les résultats de l'estimation des paramètres du modèle, si elle a déjà été effectuée. Si les résultats sont déjà affichés, le bouton suivant est affiché:  ; cliquez alors ce bouton pour cacher les résultats.

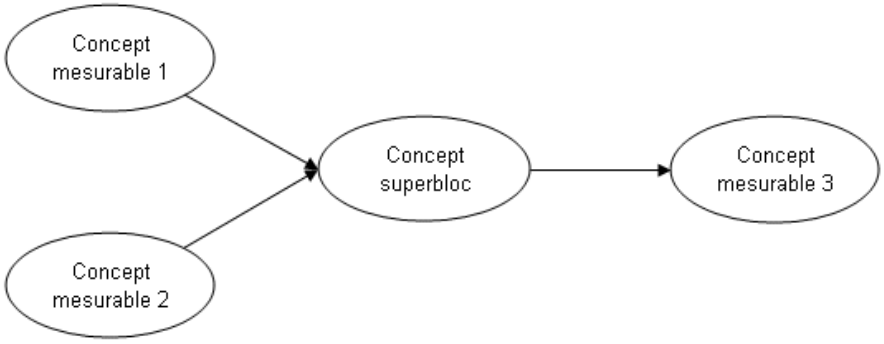
 Cliquez sur ce bouton pour afficher la boîte de dialogue des options d'affichage des résultats sur la feuille PLSPM.

 Cliquez sur ce bouton pour démarrer l'optimisation du modèle puis pour afficher les résultats dans les feuilles de résultats et sur la feuille PLSPMGraph.

Pour supprimer une flèche entre deux variables latentes, vous pouvez cliquer sur la flèche, puis utiliser Ctrl+Suppr.

Ajouter des variables manifestes

Une fois qu'une ou plusieurs variables latentes ont été ajoutées sur la feuille PLSPMGraph en utilisant la fonction appropriée de la barre d'outils "Path modeling", vous pouvez définir les variables manifestes qui correspondent à ces variables. Une variable latente est forcément liée à des variables manifestes, même dans le cas où il s'agit d'une variable superbloc. Une variable superbloc est une variable latente constituée elle-même de plusieurs variables latentes (les flèches vont des variables constitutives à la variable latente).



Pour un superbloc, l'ajout des variables manifestes est rendu très simple par l'interface de XLSTAT.

Pour ajouter des variables manifestes, vous pouvez :

- Excel 2003 et antérieures : double-cliquer sur la variable latente ou cliquer sur le bouton droit de la souris, puis choisir "Ajouter des variables manifestes".
- Excel 2007 et ultérieures : cliquer sur le bouton « MV » de la barre d'outils, ou utilisez le raccourci Ctrl+M.

Ces actions entraînent l'affichage d'une boîte de dialogue dont les options sont les suivantes :

Onglet **Général**:

Nom de la variable latente : entrez le nom de la variable latente.

Variables manifestes : sélectionnez sur la feuille D1 les données qui correspondent aux variables manifestes. Les variables peuvent être quantitatives ou qualitatives.

- **Quantitatives** : activez cette option si vous souhaitez utiliser des variables quantitatives puis sélectionnez ces variables.
- **Qualitatives** : activez cette option si vous souhaitez utiliser des variables qualitatives puis sélectionnez ces variables.

Libellés des variables : activez cette option si la première ligne des données sélectionnées comprend un en-tête.

Position : choisissez la position où les variables manifestes doivent être positionnées par rapport à la variable latente.

Mode : choisissez le mode qui détermine comment la variable latente est construite à partir des variables manifestes. Les options possibles sont "**Mode A** " (mode réflectif, les flèches sont dirigées des variables latentes vers les variables manifestes), "**Mode B** " (mode formatif, les flèches sont dirigées des variables manifestes vers les variables latentes), "**Centroïde** ", "**PCA** ", "**PLS** ", et "**Mode MIMIC** " (un mélange des Mode A et Mode B). Dans le cas du mode MIMIC, vous devez sélectionner une colonne avec une ligne par variable manifeste (et un en-tête si l'option « Libellés des variables » est activée), avec des A pour les variables en Mode A, et des B pour les variables en mode B. Pour plus de détails sur les modes, vous pouvez consulter la section description. Le mode "**Automatique** " n'est disponible que pour les superblocs. Il permet de faire en sorte que les modes des variables manifestes des variables latentes constitutives du superbloc soient réutilisés. Le mode **RGCCA** permet de saisir la valeur tau et le mode Ridge RGCCA permet d'obtenir automatiquement un tau optimal. Ces deux modes ne peuvent être appliqués qu'avec la méthode RGCCA (voir la section description).

Déflation : choisissez le mode de déflation. La déflation est utilisée lorsque le modèle est calculé sur la seconde dimension et les dimensions suivantes.

- **Pas de déflation** : quelque soit la dimension, les scores de la variable latente sont constants.
- **Externe** : Pour les dimensions successives, les résidus sont calculés à partir du modèle externe.
- **Interne** : Pour les dimensions successives, les résidus sont calculés à partir du modèle interne.
- **Interne(W)** : Pour les dimensions successives, les résidus sont calculés à partir du modèle interne après ré-estimation des poids.

Dimension : donnez le nombre de dimension que vous désirez étudier.

Inversion du signe : activez cette option si vous souhaitez changer le signe de la variable latente. Cette option est utile si vous observez que l'influence d'une variable latente est contraire à ce qu'elle devrait être.

Superbloc : vous ne pouvez activer cette option que si des variables latentes ont déjà été créées, et si des variables manifestes ont été ajoutées pour ces mêmes variables. La liste des variables latentes dont les variables manifestes ont été définies est alors ajoutée. Vous pouvez ensuite définir quelles variables latentes sont à inclure dans la variable superbloc.

Interaction : vous ne pouvez activer cette option que si des variables latentes ont déjà été créées, et si des variables manifestes ont été ajoutées pour ces mêmes variables. Une variable d'interaction est le produit de deux variables latentes qui ont la même variable successeur. La variable d'interaction aura le même successeur que les variables qui ont servi à la générer.

Onglet **Superbloc** :

La liste des variables latentes déjà construites apparaît (variables latentes génératrices). Sélectionnez les variables à inclure dans le superbloc.

Onglet **Interaction** :

- **Variables latentes génératrices** : La liste des variables latentes déjà construites apparaît. Sélectionnez deux variables latentes expliquant la variable latente à laquelle la variable interaction est reliée.
- **Transformation des données** : Sélectionnez un prétraitement pour les variables manifestes avant de faire leur produit. On peut soit les laisser dans leur échelle d'origine, soit les centrer, soit les normalisées (moyenne nulle et variance 1).

Onglet **Options (PLS)**(affichage expert) :

Options pour la régression PLS dans le modèle structurel **modèle structurel (PLS)** :

Conditions d'arrêt :

- **Automatique** : activez cette option que XLSTAT détermine automatiquement le nombre de composantes à retenir.
- **Max composantes** : activez cette option pour fixer le nombre maximum de composantes à prendre en compte dans le modèle. La valeur par défaut est 2.

Options pour la régression PLS dans le **modèle de mesure (PLS)** (actif uniquement si le mode PLS a été choisi) :

Conditions d'arrêt :

- **Automatique** : activez cette option que XLSTAT détermine automatiquement le nombre de composantes à retenir.
- **Max composantes** : activez cette option pour fixer le nombre maximum de composantes à prendre en compte dans le modèle. La valeur par défaut est 2.

Définir des groupes

Si une variable qualitative est disponible et si vous pensez qu'il pourrait y avoir des différences au niveau des valeurs des paramètres du modèle (et non de sa structure) pour les différentes catégories de cette variable, alors vous pouvez l'utiliser pour définir des groupes.

Pour définir des groupes, allez sur la feuille "PLSPMGraph", puis cliquez sur l'icône appropriée de la barre d'outils. Cela entraîne l'apparition de la boîte de dialogue des "Groupes", dont les entrées sont :

Groupes : sélectionnez sur la feuille D1 les données qui correspondent à la variable qualitative qui indique à quel groupe chaque observation appartient.

Libellé de colonne : activez cette option si la première ligne de la sélection correspond à un en-tête.

Trier alphabétiquement : activez cette option si vous voulez que XLSTAT trie alphabétiquement les noms des groupes (les modalités de la variable qualitative sélectionnée). Si cette option n'est pas activée, les modalités sont listées selon leur ordre d'apparition.

Lorsque que vous cliquez sur **OK**, une liste est ajoutée dans le coin supérieur gauche de la feuille PLSPMGraph. Une fois que le modèle a été calculé, vous pouvez utiliser cette liste pour afficher les résultats des différents groupes sur la feuille PLSPMGraph. Les résultats du modèle correspondant aux différents groupes sont aussi affichés sur des feuilles séparées.

Remarque : si vous souhaitez ne plus tenir compte de la variable de groupe, il vous suffit de cliquer sur le bouton approprié de la barre d'outils "Path modeling ".

Ajuster le modèle

Une fois le modèle conçu sur la feuille PLSPMGraph, et une fois que les variables manifestes ont été définies pour chaque variable latente, vous pouvez cliquer sur le bouton  de la barre "Path modeling" pour afficher la boîte de dialogue de définition des options pour l'ajustement du modèle.

Onglet **Général**:

Traitement des variables manifestes (affichage classique et expert) : Choisissez si et comment les variables manifestes doivent être transformées.

- **Standardisées, poids non mis à l'échelle** : les variables manifestes sont standardisées avant l'ajustement du modèle, et les poids externes correspondants sont estimés.
- **Standardisées, poids mis à l'échelle** : les variables manifestes sont standardisées avant l'ajustement du modèle, et les poids externes sont estimés pour les variables brutes.
- **Réduites, poids non mis à l'échelle** : les variables manifestes sont réduites (divisées par leur écart type) avant l'ajustement du modèle, et les poids externes correspondants sont estimés.
- **VM d'origine** : les variables manifestes ne sont pas transformées.

Poids initiaux (affichage classique et expert) : permet de choisir les valeurs initiales des poids externes au début de l'algorithme PLS.

- **Valeur du premier vecteur propre** : on utilise la valeur du premier vecteur propre comme vecteur d'initialisation.
- **Signes des coordonnées du premier vecteur propre** : au lieu de prendre la valeur, on prend le signe associé aux coordonnées du premier vecteur propre.
- **-1 pour le max du 1 ^{er} vecteur propre, sinon +1.**
- **-1 pour le min du 1 ^{er} vecteur propre, sinon +1.**
- **+1 pour la variable avec la variance max, sinon 0.**
- **-1 pour la dernière variable manifeste, sinon +1.**

Poids des observations : activez cette option si vous voulez pondérer les observations. Si vous n'activez pas cette option, les poids seront tous considérés comme valant 1. Les poids doivent être impérativement supérieurs ou égaux à 0. Si un en-tête de colonne a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

Méthode (affichage classique et expert) : Sélectionnez la méthode d'estimation. On peut choisir entre PLSPM, GSCA et RGCCA. Certaines options seront désactivées dans le cas des méthodes GSCA et RGCCA.

REBUS (affichage expert) : activez cette option si vous voulez appliquer la méthode REBUS. Lorsque vous activez sur cette option, l'onglet « REBUS » devient accessible. Pour appliquer la méthode REBUS, il faut que tous les blocs soient construits avec le mode A et qu'aucune variable de groupe ne soit sélectionnée. De plus, la méthode REBUS restreint un certain nombre d'options. Les variables manifestes doivent être standardisées ainsi que les scores des variables latentes. D'autre part, les méthodes de traitement des données manquantes NIPALS et Lohmöller ne sont plus utilisables.

RGCCA (affichage expert) : activez cette option si vous voulez appliquer la méthode RGCCA. Il faut que tous les blocs soit définis soit avec le mode A, soit avec le mode B, soit avec le mode RGCCA.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des variables : activez cette option si la première ligne des données sélectionnées contient un libellé.

Libellés des observations : activez cette option si vous voulez utiliser des libellés d'observations disponibles sur une feuille Excel pour l'affichage des résultats. Si l'option « Libellés des variables » est activée, la première cellule de la sélection doit comprendre un en-tête. Si vous n'activez pas cette option, des libellés seront automatiquement créés (Obs1, Obs2, …).

Echelle des variables manifestes (affichage Market) : permet de choisir l'échelle dans laquelle les variables manifestes sont présentées. Deux options sont possibles :

- Automatique : dans ce cas les variables manifestes sont standardisées.
- Choisir : dans ce cas, on peut entrer le minimum et le maximum des variables manifestes (échelle uniforme) afin de travailler sur les variables d'origine.

Echelle des variables latentes (affichage Market) : permet de choisir l'échelle des scores des variables latentes. Deux options sont possibles :

- Echelle 0-100 : les scores des variables latentes sont donnés sur une échelle entre 0 et 100.
- Echelle des variables manifestes : les scores des variables latentes sont donnés dans la même échelle que les variables manifestes.

Onglet **Options** :

Estimation interne (affichage classique et expert) : choisissez la méthode d'estimation du modèle interne (voir la section description pour plus de détails).

- **Structurel** : les poids internes sont égaux à la corrélation entre les variables latentes lorsque l'on estime une variable latente explicative (prédécesseur). Sinon ils sont égaux aux coefficients de la régression OLS.
- **Factoriel** : les poids internes sont égaux à la corrélation entre les variables latentes.
- **Centroïde** : les poids internes sont égaux au signe de la corrélation entre les variables.
- **PLS** : les poids internes sont égaux à la corrélation entre les variables latentes lorsque l'on estime une variable latente explicative (prédécesseur). Sinon ils sont égaux aux coefficients de la régression PLS.

Régression (affichage expert) : choisissez la méthode de régression pour l'estimation des « path coefficients » :

- **OLS** : régression par les moindres carrés.
- **PLS** : régression par les moindres carrés partiels.

Dimensions (affichage expert) : entrez le nombre de dimensions jusqu'auquel le modèle doit être calculé.

Conditions d'arrêt :

- **Itérations** : entrez le nombre maximal d'itérations pour l'algorithme. Les calculs sont interrompus dès que le nombre maximal d'itérations est dépassé. Valeur par défaut : 100.
- **Convergence** : entrez la valeur seuil d'évolution maximale des communalités d'une itération à l'autre, qui une fois atteinte permet de considérer que l'algorithme a convergé. Valeur par défaut : 0,0001.

Intervalles de confiance : activez cette option pour calculer les intervalles de confiance. Choisissez ensuite la méthode à utiliser pour calculer les intervalles :

- **Bootstrap** : activez cette option pour utiliser la méthode bootstrap. Entrez ensuite le nombre de rééchantillonnages générés pour calculer les intervalles de confiance.
- **Jackknife** : activez cette option pour utiliser la méthode jackknife. Entrez ensuite la taille des groupes générés pour calculer les intervalles de confiance.

Intervalle de confiance (%) : entrez la taille en % des intervalles de confiance.

Estimations rééchantillonnées (affichage expert) : activez cette option pour afficher les valeurs des corrélations du modèle externe et des coefficients structurels pour chaque

échantillon généré. De plus, les écart- types et les bornes des intervalles de confiance associés aux effets indirects sont aussi affichés.

Qualité du modèle (classique et affichage expert) :

- **Blindfolding** : activez cette option pour évaluer la qualité du modèle en utilisant l'approche « blindfolding » (voir la section description pour plus de détails). Des valeurs de validation croisée seront alors calculées pour la redondance et les communalités.

Segmentation (affichage market) : activez cette option pour rechercher des groupes homogènes d'observations par rapport au modèle spécifié. Il s'agit de l'application de l'algorithme REBUS. On peut ensuite sélectionner la troncature à effectuer.

Comparaisons (affichage market) : si des groupes ont été sélectionnés les tests t et de permutation peuvent être activés à ce niveau.

Onglet **Données manquantes** :

Ne pas accepter les valeurs manquantes : activez cette option pour que XLSTAT empêche la poursuite des calculs si des valeurs manquantes sont détectées.

Supprimer les observations : activez cette option pour supprimer les observations comportant des données manquantes.

Utiliser NIPALS : activez cette option pour utiliser l'algorithme NIPALS pour la gestion des données manquantes (voir la section description pour plus de détails).

Lohmöller : activez cette option pour utiliser la procédure de Lohmöller pour la gestion des données manquantes (si la régression PLS est utilisée à la place de la régression OLS, alors le modèle est appliqué sur les données disponibles) :

- **Utiliser la moyenne ipsative** : activez cette option pour utiliser la moyenne des variables latentes pour déterminer les données manquantes au niveau des variables manifestes.
- **Renormaliser** : activez cette option pour renormaliser les poids externes au niveau de chaque observation lorsqu'il y a des données manquantes.

Remarque: dans le cas de poids standardisés, les deux options ci-dessus reviennent à une suppression par paires pour calculer les moyennes et les écarts-types, et à une imputation par la moyenne pour le calcul des scores.

Estimer les données manquantes : activez cette option pour estimer les données manquantes avant le début des calculs.

- **Moyenne ou mode** : activez cette option pour estimer les données manquantes en utilisant la moyenne (variables quantitatives) ou le mode (variables qualitatives) pour les variables correspondantes.

- **Plus proche voisin** : activez cette option pour estimer les données manquantes d'une observation en recherchant le plus proche voisin de l'observation.

Onglet **Tests multigroupes**(affichage expert) :

Si le nombre de groupes sélectionné est plus grand que 2, alors cet onglet apparaît.

Test t multigroupes : activez cette option pour tester l'égalité des coefficients structurels du modèle entre les groupes (le nombre d'échantillons bootstrap utilisé est défini dans l'onglet options).

- **Niveau de signification (%)** : entrez le niveau de signification pour les tests t multigroupes.

Test de permutation : activez cette option pour tester l'égalité des paramètres du modèle entre les groupes (le nombre de groupes est limité à 2).

- **Nombre de permutations** : entrez le nombre de permutations désirées.
- **Niveau de signification (%)** : entrez le niveau de signification pour les tests de permutation.
- **Path coefficients** : activez cette option pour tester l'égalité des path coefficients.
- **Corrélations** : activez cette option pour tester l'égalité des corrélations entre variables manifestes et variables latentes.
- **Qualité du modèle** : activez cette option pour tester l'égalité des indice de qualité du modèle (communalités, redondances et GoF).

Onglet **REBUS**(affichage expert) :

Cet onglet est affiché si l'option REBUS est activée.

Troncature :

- **Automatique** : activez cette option pour que le nombre de classes soit défini automatiquement à l'intérieur de l'algorithme.
- **Nombre de classes** : activez cette option si vous désirez entrer manuellement le nombre de classes.

Conditions d'arrêt :

- **Itérations** : entrez le nombre maximal d'itérations pour l'algorithme REBUS. Les calculs sont interrompus dès que le nombre maximal d'itérations est dépassé. Valeur par défaut : 100.
- **Seuil (%)** : entrez la valeur seuil afin de considérer que les classes sont stables. Valeur par défaut : 95.

Dendrogramme :

- **Horizontal** : choisissez cette option pour afficher un dendrogramme horizontal.
- **Vertical** : choisissez cette option pour afficher un dendrogramme vertical.
- **Etiquettes** : activez cette option pour afficher les libellés des objets (dendrogramme complet) ou des classes (dendrogramme tronqué) sur le dendrogramme.
- **Couleurs** : activez cette option pour utiliser des couleurs pour représenter les différents groupes sur le dendrogramme complet.

Onglet **Sorties**:

Statistiques descriptives : activez cette option pour afficher les statistiques descriptives pour les variables sélectionnées.

Modèle : activez cette option pour afficher les spécifications du modèle.

Corrélations (affichage classique et expert) : activez cette option pour afficher la matrice de corrélations ou de covariance en fonction du type d'options choisi dans l'onglet « Général ».

- **Tester la significativité** : dans le cas où une corrélation a été choisie dans l'ongle « Général » de la boîte de dialogue, activez cette option pour tester la significativité des corrélations.
- **Niveau de signification (%)** : entrez le niveau de signification pour les tests ci-dessus.

Détéction des valeurs extrêmes (affichage expert) : Activez cette option pour afficher les tableaux DModX et DmodY dans le cas de la régression PLS.

VM après déflation (affichage expert) : Activez cette option pour afficher les variables manifestes après déflation lorsque plus d'une dimension a été sélectionnée.

Corrélations Variables/Facteurs : activez cette option pour afficher les corrélations entre les facteurs et les variables.

Modèle interne : activez cette option pour afficher les résultats qui correspondent au modèle interne.

Modèle externe : activez cette option pour afficher les résultats qui correspondent au modèle externe.

R^2 et communalités : activez cette option pour afficher les R2 des variables latentes du modèle structurel et les communalités des variables manifestes.

Qualité du modèle : activez cette option pour afficher les résultats de la procédure blindfolding.

Scores des variables latentes :

- **Standardisés** : activez cette option pour calculer et afficher les scores standardisés.

- **Utilisant les poids normalisés** : activez cette option pour afficher les scores calculés avec des poids normalisés.
- **Standardisés > 0-100** : activez cette option pour calculer les scores standardisés, et pour ensuite les transformer et les afficher sur une échelle 0-100.
- **Utilisant les poids normalisés > 0-100** : activez cette option pour calculer les scores factor scores en utilisant les poids normalisés, puis pour transformer et afficher les scores sur une échelle 0-100.

Simulations (affichage market) : activez cette option afin d'afficher les tableaux de simulation permettant de visualiser l'impact sur une variable latente cible d'une modification d'une variable manifeste ou latente.

- Variable latente cible : sélectionnez la variable latente à expliquer. Il faut sélectionner une variable latente endogène.
- Echelle des modifications : sélectionnez l'échelle des modifications, il peut soit s'agir d'un nombre de point, soit d'un pourcentage. Une fois cette option sélectionné, vous pouvez entrer le minimum, le maximum et le pas de changement pour obtenir la plage des valeurs à tester.

Tableaux IPMA (affichage market) : activez cette option si vous désirez afficher les tableaux basés sur l'IPMA (Importance Perform Analysis).

Onglet **Graphiques** :

Graphique des coefficients : activez cette option pour afficher les coefficients normalisés du modèle interne.

Graphique IPMA : activez cette option pour afficher les graphiques IPMA.

Graphique de simulation (variables manifestes) (affichage market) : activez cette option afin d'afficher les graphiques de simulation pour l'impact d'une modification des variables manifestes sur le score de la variable latente cible.

Graphique de simulation (variables latentes) (affichage market) : activez cette option afin d'afficher les graphiques de simulation pour l'impact d'une modification des variables latentes sur le score de la variable latente cible.

Options pour les résultats

De nombreux résultats peuvent être affichés sur la feuille PLSPMGraph, une fois que le modèle a été ajusté. Pour afficher la boîte des options correspondante, cliquez sur l'icône  de la barre "Path modeling".

Onglet **Variables latentes** :

Ces options permettent de définir quels résultats sont affichés sous les variables latentes.

- **Moyenne :** activez cette option pour afficher la moyenne de la variable latente.
- **Moyenne (Bootstrap):** activez cette option pour afficher la moyenne de la variable latente, calculée en utilisant une procédure bootstrap.
- **Intervalle de confiance :** activez cette option pour afficher l'intervalle de confiance autour de la moyenne.
- **R^2 :** activez cette option pour afficher le coefficient de détermination R^2 entre les variables manifestes et la variable latente.
- **R^2 ajusté :** activez cette option pour afficher le coefficient de détermination R^2 ajusté entre les variables manifestes et la variable latente.
- **R^2 (Boot/Jack):** activez cette option pour afficher le coefficient de détermination R^2 entre les variables manifestes et la variable latente, calculé en utilisant une procédure bootstrap ou jackknife.
- **R^2 (int. de conf.):** activez cette option pour afficher l'intervalle de confiance du coefficient de détermination R^2 entre les variables manifestes et la variable latente, calculé en utilisant une procédure bootstrap ou jackknife.
- **Communalité :** activez cette option pour afficher la communalité entre la variable latente et les variables manifestes.
- **Redondance :** activez cette option pour afficher la redondance entre la variable latente et les variables manifestes.
- **Communalité (Blindfolding):** activez cette option pour afficher la communalité entre la variable latente et les variables manifestes, calculée en utilisant la procédure de blindfolding.
- **Redondance (Blindfolding):** activez cette option pour afficher la redondance entre la variable latente et les variables manifestes, calculée en utilisant la procédure de blindfolding.
- **rho de D.G.:** activez cette option pour afficher le coefficient rho de Dillon-Goldstein.
- **alpha de Cronbach:** activez cette option pour afficher l'alpha de Cronbach.

- **Ecart-type (scores) :** activez cette option pour afficher les écart-types associés aux coordonnées.

Onglet **Flèches (variables latentes)** :

Ces options permettent de définir quels résultats sont affichés sur les flèches qui relient les variables latentes.

- **Corrélation :** activez cette option pour afficher le coefficient de corrélation coefficient entre les deux variables latentes.
- **Contribution :** activez cette option pour afficher la contribution des variables latentes au R^2.
- **Path coefficient :** activez cette option pour afficher le coefficient de régression qui correspond à la régression de la variable latente qui se trouve à la pointe de la flèche (variable dépendante) par les variables latentes qui se trouvent à l'origine de la flèche (variable prédécesseur ou explicative).
- **Path coefficient (B/J) :** activez cette option pour afficher le coefficient de régression qui correspond à la régression de la variable latente qui se trouve à la pointe de la flèche (variable dépendante) par les variables latentes qui se trouvent à l'origine de la flèche (variable prédécesseur ou explicative), calculé en utilisant une méthode bootstrap ou jackknife.
- **Ecart-type :** activez cette option pour afficher l'écart-type correspondant au coefficient de régression.
- **Intervalle de confiance :** activez cette option pour afficher l'intervalle de confiance correspondant au coefficient de régression.
- **Coeff. norm.:** activez cette option pour afficher le coefficient de régression normalisé.
- **t de Student :** activez cette option pour afficher la valeur du t de Student.
- **Pr > |t| :** activez cette option pour afficher la p-value qui correspond au t de Student.
- **Corrélations partielles :** activez cette option pour afficher les corrélations partielles entre les variables latentes.
- **L'épaisseur des flèches dépend de :** L'épaisseur des flèches peut être liée à :
- La p-value associée au t de Student (plus la valeur est faible, plus la flèche est épaisse).
- La corrélation (plus son carré est élevé, plus les flèches sont épaisses; une flèche bleue correspond à une corrélation négative, une flèche rouge à une corrélation positive).
- La contribution (plus la valeur est élevée, plus la flèche est épaisse).

Onglet **Flèches (Variables manifestes)** :

Ces options permettent de définir quels résultats sont affichés sur les flèches qui relient les variables manifestes à leur variable latente.

- **Poids :** activez cette option pour afficher le poids.
- **Poids (Bootstrap):** activez cette option pour afficher le poids calculé avec une méthode bootstrap.
- **Poids normalisé:** activez cette option pour afficher le poids normalisé.
- **Ecart-type :** activez cette option pour afficher l'écart-type du poids.
- **Intervalle de confiance:** activez cette option pour afficher l'intervalle de confiance sur le poids.
- **Corrélation :** activez cette option pour afficher le coefficient de corrélation entre la variable manifeste et la variable latente.
- **Corrélation (Boot/Jack) :** activez cette option pour afficher le coefficient de corrélation entre la variable manifeste la variables latente, calculée en utilisant une procédure bootstrap ou jackknife.
- **Corrélation (écart-type) :** activez cette option pour afficher l'écart-type du coefficient de corrélation entre la variable manifeste et la variable latente, calculé en utilisant une procédure bootstrap ou jackknife.
- **Corrélation (intervalle de confiance) :** activez cette option pour afficher l'intervalle de confiance pour le coefficient de corrélation entre la variable manifeste et la variable latente, calculé en utilisant une procédure bootstrap ou jackknife.
- **Communalités :** activez cette option pour afficher la communalité entre la variable latente et la variable manifeste.
- **Redondance :** activez cette option pour afficher la redondance entre la variable latente et la variable manifeste.
- **Communalité (Blindfolding) :** activez cette option pour afficher la communalité entre la variable latente et la variable manifeste, calculée en utilisant la procédure de blindfolding.
- **Redondance (Blindfolding) :** activez cette option pour afficher la redondance entre la variable latente et la variable manifeste, calculée en utilisant la procédure de blindfolding.
- **L'épaisseur des flèches dépend de :** L'épaisseur des flèches peut être liée à :
- La corrélation (plus son carré est élevé, plus les flèches sont épaisses; une flèche bleue correspond à une corrélation négative, une flèche rouge à une corrélation positive).
- Poids normalisés.

Résultats

Les premiers résultats obtenus sont calculés avant l'application de l'approche PLS :

Statistiques simples : Ce tableau rassemble pour toutes les variables manifestes, le nombre d'observations, le nombre de données manquantes, le nombre de données non manquantes, le minimum, le maximum, la moyenne et l'écart-type.

Spécification du modèle (modèle de mesure) : Ce tableau rassemble pour chaque variable latente, le nombre de variables manifestes associées, le mode, le type de variable (endogène ou exogène), le fait d'avoir inverser son signe, le nombre de dimensions estimées et la liste des variables manifestes qui lui sont associées.

Spécification du modèle (modèle structurel) : Cette matrice carrée permet de voir s'il existe une flèche allant de la variable en colonne à la variable en ligne.

Matrice de corrélation : Si l'option a été sélectionnée, la matrice de corrélation entre les variables manifestes apparaît.

Fiabilité du bloc (Composite reliability) : Ce tableau permet de vérifier l'unidimensionnalité des blocs de variables manifestes. Pour chaque variable latente, une analyse en composantes principales est effectuée sur la matrice de covariance ou de corrélation des variables manifestes afin d'estimer le nombre de dimension significatives. L'alpha de Cronbach, le rho de Dillon et Goldstein, la valeur critique pour les valeurs propres (que l'on peut comparer à la valeur obtenue par l'ACP) et le nombre de conditionnement sont calculés afin de déterminer le nombre de dimensions à traiter.

Corrélations Variables/Facteur (Variable latente X / Dimension Y) : Ces tableaux affichent pour chaque variable latente et pour chaque dimension, la corrélation entre les variables manifestes et le facteur obtenus lors de l'application d'une ACP.

Les résultats suivants sont obtenus après application de l'approche PLS.

Qualité de l'ajustement (GoF) (Dimension Y) : Ce tableau rassemble les indices de qualité d'ajustement issus du GoF. Quatre variantes de cet indice sont détaillées :

- **Absolu :** Valeur du GoF.
- **Relatif :** GoF relative obtenu en divisant le GoF absolu par sa valeur maximale sur le jeu de données étudié.
- **Modèle externe :** Composante du GoF basée sur la performance du modèle de mesure (associé aux communalités).
- **Modèle interne :** Composante du GoF basée sur la performance du modèle structurel (associé aux R2 des variables latentes endogènes).

Cross-loadings (variables manifestes monofactorielles / dimension Y) : Ce tableau permet de vérifier le fait qu'une variable manifeste est réellement monofactorielle. Si le modèle a été

bien spécifié, les variables manifestes doivent être en forte relation avec la variable latente qui leur est associée.

Modèle externe (dimension Y) :

- **Poids (dimension Y) :** Coefficient associé à chaque variable manifeste dans la combinaison linéaire utilisée pour calculer le score de la variable latente.
- **Corrélations (dimension Y) :** Corrélation entre chaque variable manifeste et la variable latente qui lui est associée. Dans ce tableau, on trouve aussi les loadings ainsi que la location associée.

Modèle externe (dimension Y) :

- **R^2 (variable latente X / dimension Y) :** Valeur du R^2 pour les variables latentes endogènes.
- **Path coefficient (variable latente X / dimension Y) :** Valeur du coefficient de régression du modèle structurel estimé à partir des scores standardisés des variables latentes. Dans ce tableau, on trouve aussi le coefficient f^2 (size effect).
- **Impact et contribution des variables pour variable latente X (dimension Y) :** Valeur des corrélations, des path coefficients, du produit des corrélations et des path coefficients et contribution au R^2 des variables latentes expliquant la variable latente X. Un graphique est associé à ce résultat.

Evaluation du modèle (dimension Y) : Le tableau rassemble les résultats importants associés aux scores des variables latentes.

Corrélation (Variable latente) / dimension Y (affichage expert) : Matrice de corrélations obtenue à partir des scores des variables latentes.

Corrélations partielles (Variable latente) / dimension Y (affichage expert): Matrice des corrélations partielles entre les scores des variables latentes.

Effets directs (Variable latente) / dimension Y (affichage expert) : Matrice de corrélations illustrant l'effet des variables latentes expliquant directement une autre variable latente (lorsque les variables latentes ne sont pas reliées directement alors cette corrélation est nulle).

Effets indirects (Variable latente) / dimension Y (affichage expert) : Matrice de corrélations illustrant l'effet des variables latentes n'expliquant pas directement une autre variable latente (lorsque les variables latentes sont reliées directement alors cette corrélation est nulle). L'information passe par d'autres variables latentes avant d'arriver à la variable latente étudiée. Si l'option estimations rééchantillonnées a été sélectionnée, les écart-types et les bornes des intervalles de confiance obtenus par rééchantillonnage sont affichés dans trois tableaux distincts.

Effets totaux (Variable latente) / dimension Y (affichage expert) : Matrice des corrélations totales. Effets totaux = Effets directs + Effets indirects.

Validité discriminante (Corrélations carrées < AVE) (Dimension Y) : Ce tableau permet de vérifier si chaque variable latente peut être associée à un concept distinct de celui associé aux autres variables latentes du modèle. Le R^2 associé à chaque pair de variables latentes doit être

plus petit que la communalité moyenne. Ceci montre que une plus grande part de variance est partagée entre chaque variable latente et son bloc qu'entre deux variables latentes différentes.

Tableaux et graphiques IPMA (Importance Performance Matrix Analysis) (affichage expert et market) : Ces tableaux rassemblent pour chaque variable latente endogène les valeurs des importances et des performances des variables latentes. L'importance est égale à l'effet total sur la variable latente endogène étudiée alors que la performance est le score de la variable latente ramenée sur une échelle entre 0 et 100. Ces indices sont représentés sur des graphiques.

Tableaux et graphiques de simulation (affichage market) : Cet ensemble de résultats permet de comprendre l'impact d'un changement d'une variable du modèle sur une variable latente cible.

- Le premier tableau rassemble les variables latentes les plus importantes pour la prédiction de la variable cible.
- Le second tableau rassemble les variables manifestes les plus importantes pour la prédiction de la variable cible.
- Le tableau et le graphique suivant permettent de visualiser l'impact d'un changement d'une variable manifeste sur la variable latente cible.
- Le tableau et le graphique suivant permettent de visualiser l'impact d'un changement d'une variable manifeste sur la moyenne du score de la variable latente cible (il s'agit ici de moyenne et plus de changement).
- Le tableau et le graphique suivant permettent de visualiser l'impact d'un changement d'une variable latente sur la variable latente cible.
- Le tableau et le graphique suivant permettent de visualiser l'impact d'un changement d'une variable latente sur la moyenne du score de la variable latente cible (il s'agit ici de moyenne et plus de changement).

Scores des variables latentes (dimension Y) :

- Moyenne / Scores des variables latentes (Dimension Y): Valeur moyenne des scores des variables latentes.
- Statistiques simples / Scores des variables latentes (dimension Y) : Statistiques descriptives sur les scores des variables latentes obtenues à partir du modèle de mesure.
- Scores des variables latentes (dimension Y) : Scores des variables latentes au niveau de chaque individu obtenus par une combinaison linéaire des variables manifestes.
- Statistiques simples / Scores prédits avec le modèle structurel (dimension Y) : Statistiques descriptives des scores des variables latentes obtenues à partir du modèle structurel.
- Scores prédits avec le modèle structurel (dimension Y) : Scores des variables latentes en se basant sur le modèle structurel.

Bootstrap : Si l'option estimations rééchantillonées a été sélectionnée, les valeurs des corrélations du modèle externe et les coefficients structurels pour chaque échantillon bootstrap sont affichés.

Evaluation du modèle / Modèle externe (Blindfolding / Dimension Y) : Résultats des cv-communalités obtenues par la procédure blindfolding (pour plus de détails voir la partie description).

Evaluation du modèle / Modèle interne (Blindfolding / Dimension Y) : Résultats des cv-redondances obtenues par la procédure blindfolding (pour plus de détails voir la partie description).

Si plusieurs groupes ont été définis, d'autres sorties peuvent apparaître :

Feuille PLSPM(Groupe) : Pour chaque groupe, une feuille de résultats apparaît avec tous les résultats précédents associés à chacun des groupes.

Feuille PLSPM(Test t multigroup) : Pour chaque path coefficient, un test est effectué entre toutes les paires de groupes.

- Différence : différence en valeur absolue observé entre les coefficients d'un groupe à l'autre.
- t (valeur observée) : valeur observé de la statistique t.
- t (valeur critique) : valeur critique pour la statistique t.
- DDL : nombre de degrés de liberté.
- p-value : p-valeur associée au test t.
- Alpha : Niveau de signification
- Significatif : si oui la différence entre les path coefficients est significative, si non, elle ne l'est pas.

Feuille PLSPM (Test de permutation) : Pour chaque paramètre sélectionné, un test basé sur des permutations est effectué entre les deux groupes.

- Différence : différence en valeur absolue observé entre les paramètres d'un groupe à l'autre.
- p-value : p-valeur associée au test de permutation.
- Alpha : Niveau de signification
- Significatif : si oui la différence entre les paramètres est significative, si non, elle ne l'est pas.

Si l'option REBUS a été activée, d'autres sorties peuvent apparaître :

Feuille REBUS : le dendogramme issu de la CAH est présenté. Les classes obtenues avec l'algorithme REBUS et les indices (CM index).

Feuille PLSPM(Classe) : pour chaque classe, une feuille de résultats apparaît avec tous les résultats précédents associés à chacune des classes.

Exemple

Des tutoriels sur l'utilisation du module XLSTAT-PLSPM sont disponibles sur le Centre d'aide XLSTAT :

Avec Excel 2007 et 2010 :

http://www.xlstat.com/demo-plspm2007f.htm

Avec Excel 2003 ou antérieur :

http://www.xlstat.com/demo-plspmf.htm

Un tutoriel montrant comment comparer plusieurs groupes d'observations avec XLSTAT-PLSPM est disponible sur le Centre d'aide XLSTAT :

http://www.xlstat.com/demo-plspmgrpf.htm

Un tutoriel montrant comment appliquer l'algorithme REBUS pour déterminer des classes d'observations à partir d'un modèle PLS avec XLSTAT est disponible sur leCentre d'aide XLSTAT :

http://www.xlstat.com/demo-plspmREBUSf.htm

Bibliographie

Amato S., Esposito Vinzi V. and Tenenhaus M. (2004). A global Goodness- of-Fit index for PLS structural equation modeling. in: Proceedings of the XLII SIS Scientific Meeting, vol. Contributed Papers, 739-742, CLEUP, Padova, 2004.

Carroll J.D. (1968). A generalization of Canonical Correlation Analysis to three or more sets of variables. *Proc. 76th Conv. Am. Psych. Assoc.*, 227-228.

Chin W.W. (1998). The Partial Least Squares approach for structural equation modeling. In: G.A. Marcoulides (Ed.), Modern Methods for Business Research, Lawrence Erlbaum Associates, 295-336.

Chin W. and Dibbern J. (2010). An Introduction to a Permutation Based Procedure for Multi-Group PLS Analysis: Results of Tests of Differences on Simulated Data and a Cross Cultural Analysis of the Sourcing of Information System Services between Germany and the USA . Handbook of Partial Least Squares, Springer, 171-195.

de Leeuw, J., Young, F. W., & Takane, Y. (1976). Additive structure in qualitative data: An alternating least squares method with optimal scaling features. *Psychometrika*, ***41***, 471–503.

Escofier B. and Pagès J. (1994). Multiple Factor Analysis, (AFMULT Package). *Computational Statistics and Data Analysis*, **18**, 121-140.

Esposito Vinzi V., Chin W., Henseler J. and Wang H. (2010). Handbook of Partial Least Squares: Concepts, Methods and Applications, Springer-Verlag.

Esposito Vinzi V., Trinchera L., Squillacciotti S. and Tenenhaus M. (2008). REBUS-PLS: A response-based procedure for detecting unit segments in PLS path modelling. Appl. Stochastic Models Bus. Ind., 24, 439–458.**Fornell C. and Cha J. (1994)**. Partial Least Squares. In: R.P. Bagozzi (Ed.), Advanced Methods of Marketing Research, Basil Blackwell, Cambridge, Ma., 52-78.

Guinot C., Latreille J. and Tenenhaus M. (2001). PLS Path Modelling and Multiple Table Analysis. Application to the cosmetic habits of women in Ile- de-France. *Chemometrics and Intelligent Laboratory Systems*, **58**, 247-259.

Horst P. (1961). Relations among M sets of variables. *Psychometrika*, **26**, 126-149.

Horst P. (1965). Factor Analysis of data matrices. Holt, Rinehart and Winston, New York.

Hwang, H., and Takane, Y. (2004). Generalized structured component analysis. *Psychometrika*, **69**, 81-99.

Jöreskog K.G. (1970). A General Method for Analysis of Covariance Structure. *Biometrika*, **57**, 239-251.

Jöreskog, K.G. and Wold, H. (1982). The ML and PLS Techniques for Modeling with Latent Variables: Historical and Comparative Aspects. In: K.G. Jöreskog and H. Wold (Eds.), Systems Under Indirect Observation, Part 1, North-Holland, Amsterdam, 263-270.

Lohmöller J.-B. (1989). Latent Variables Path Modeling with Partial Least Squares. Physica-Verlag, Heildelberg.

Pagès J. and Tenenhaus, M. (2001). Multiple Factor Analysis combined with PLS Path Modelling. Application to the analysis of relationships between physicochemical variables, sensory profiles and hedonic judgements. *Chemometrics and Intelligent Laboratory Systems*, **58**, 261-273.

Tenenhaus M. (1998). La Régression PLS. Éditions Technip, Paris.

Tenenhaus M. (1999). L'approche PLS. *Revue de Statistique Appliquée*, **47(2)**, 5-40.

Tenenhaus M., Esposito Vinzi V., Chatelin Y.-M. and Lauro C. (2005). PLS Path Modeling. *Computational Statistics & Data Analysis*, **48(1)**, 159-205.

Tenenhaus M. and Hanafi M. (2007). A bridge between PLS path modeling and multi-block data analysis. In: Esposito Vinzi V.et al. (Eds.), Handbook of Partial Least Squares: Concepts, Methods and Applications, Springer-Verlag.

Tenenhaus M. and Tenenhaus A. (2011). Regularized Generalized Canonical Correlation Analysis, Psychometrika, **76(2)**, 257-284

Wold H. (1966). Estimation of Principal Components and Related Models by Iterative Least Squares. In: P.R. Krishnaiah (Ed.), Multivariate Analysis, Academic Press, New York, 391-420.

Wold H. (1973). Non-linear Iterative PArtial Least Squares (NIPALS) modelling. Some current developments. In: P.R. Krishnaiah (Ed.), Multivariate Analysis III, Academic Press, New York, 383-407.

Wold H. (1975). Soft Modelling by latent variables: the Non-linear Iterative PArtial Least Squares (NIPALS) Approach. In: J. Gani (Ed.), Perspectives in Probability and Statistics: Papers, in Honour of M.S. Bartlett on the occasion of his sixty-fifth birthday, Applied Probability Trust, Academic, London, 117-142.

Wold H. (1979). Model construction and evaluation when theoretical knowledge is scarce: an example of the use of Partial Least Squres. Cahier 79.06 du Département d'économétrie, Faculté des Sciences Économiques et Sociales. Genève: Université De Genève.

Wold H. (1982). Soft Modeling: The basic design and some extensions. In: K.G. Jöreskog and H. Wold (Eds.), Systems under indirect observation, Part 2, North-Holland, Amsterdam, 1-54.

Wold H. (1985). Partial Least Squares. In: S. Kotz and N.L. Johnson (Eds.), Encyclopedia of Statistical Sciences, John Wiley & Sons, New York, 6, 581-591.

XLSTAT-LG

Classification par les classes latentes

Cet outil permet de classifier des cas (observations, individus) au sein de groupes (classes latentes) différents selon un ou plusieurs paramètres, dans le cadre de modèles de classification par les classes latentes (CL). Les modèles de classification CL effectuent des classifications selon des combinaisons de variables continues et/ou catégorielles (nominales ou ordinales).

Dans cette section :

Description

XLSTAT-LG est un outil de classification puissant basé sur deux modules de Latent GOLD® 5.0 : **modèles de classification par les classes latentes** et **modèles de régression sur classes latentes**.

Les analyses en classes latentes (ACL) impliquent la construction de classes latentes (CL), qui sont des sous-groupes ou segments non-observés (latents) de cas (observations, individus). Les CL sont construites en se basant sur les réponses observées (manifestes) des cas sur un ensemble de variables indicatrices. Les cas se trouvant dans la même classe latente sont similaires sur le plan de leurs réponses tandis que les cas se trouvant dans des classes différentes le sont moins. Formellement, les classes sont représentées par K catégories distinctes d'une variable nominale latente X. Comme X est catégorielle, la modélisation LC se distingue d'approches plus traditionnelles telles que l'analyse factorielle, les modèles d'équations structurelles, ainsi que les modèles de régression impliquant des effets aléatoires. Ces approches sont plutôt basées sur des variables latentes continues.

XLSTAT-LG comprend deux modules pour l'estimation de deux structures de modèles : **modèles de classification CL** et **modèles de régression CL**. Ces deux modules sont utilisables dans des domaines plus ou moins différents. Dans la suite de l'aide, le terme « segments » se réfère à « classes latentes ».

Le modèle de classification CL :

- Comprend une variable latente X à K catégories, chaque catégorie représentant une classe.
- Chaque classe comprend un groupe homogène d'individus (cas) partageant les mêmes intérêts, valeurs, caractéristiques et/ou comportements (en d'autres termes, qui partagent des paramètres de modèle communs).
- Ces intérêts, valeurs, caractéristiques et/ou comportements sont les variables observées (indicateurs) Y à partir desquelles les CL sont construites.

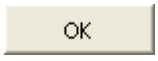
Parmi les avantages de la méthode par rapport à des approches traditionnelles de classification, citons la présence de critères de sélection de modèles et des classifications probabilistes. Les probabilités d'appartenance a posteriori sont estimées directement à partir des paramètres du modèle et sont utilisées pour assigner chaque cas à la classe modale correspondante, à savoir la classe associée à la probabilité d'appartenance a posteriori la plus élevée.

Equation de scoring : ces modèles fournissent une équation de scoring permettant de calculer les probabilités d'appartenance a posteriori directement à partir de variables observées (indicateurs). Cette équation peut être utilisée pour affilier de nouveaux cas à la classe la plus vraisemblable. Cette fonctionnalité est exclusive des modèles de classification CL.

L'équation de scoring est obtenue sous forme de cas spécial de la troisième étape dans la construction des modèles de classification CL (Vermunt 2010). La première étape implique une estimation des paramètres du modèle. Au cours de la deuxième étape, les cas sont assignés à des classes selon leurs probabilités d'appartenance a posteriori. Dans le cadre de la troisième étape, les classes latentes sont utilisées en tant que prédicteurs ou variables dépendantes pour des analyses ultérieures. Pour plus de détails, cf. Vermunt et Magidson (2013).

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.

Onglet **Général** :

Tableau des observations/variables :

Continues : Sélectionnez les variables quantitatives continues. Les données doivent être quantitatives continues. Si le libellé des variables a été sélectionné, veuillez vérifier que l'option « Libellés des colonnes » est activée.

Nominales : Sélectionnez les variables nominales (qualitatives). Les données doivent être nominales. Si le libellé des variables a été sélectionné, veuillez vérifier que l'option « Libellés des colonnes » est activée.

Ordinales : Sélectionnez les variables ordinales. Les données doivent être numériques. Si le libellé des variables a été sélectionné, veuillez vérifier que l'option « Libellés des colonnes » est activée.

Effets directs : Activez cette option si vous souhaitez indiquer un effet direct dans le modèle. Après avoir cliqué sur « OK » ceci aura pour conséquence l'apparition d'une boîte contenant toutes les paires de variables éligibles pour un paramètre d'effet direct. Afin d'inclure un effet direct, cochez la case correspondante. Des paramètres associés à cet effet seront ainsi estimés. L'inclusion d'effets directs est une manière d'atténuer l'hypothèse de dépendance locale.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des colonnes : activez cette option si la première ligne des données sélectionnées (variables dépendantes et explicatives) contient des libellés.

Libellés des observations : Activez cette option si des libellés sont disponibles pour les N observations. Puis sélectionnez les données correspondantes. Si l'option « libellés des colonnes » a été activée, la première cellule de la sélection doit comprendre un en-tête.

Pour les mesures répétées (plusieurs mesures par cas), les libellés des observations servent d'identifiants qui regrouperont ensemble les mesures pour chaque cas. Si vous n'activez pas cette option, des libellés seront automatiquement créés (Obs1, Obs2, …) et il y aura autant de cas que d'observations.

Poids des observations : Activez cette option is vous souhaitez attribuer un poids aux observations. Si vous n'activez pas cette option, tous les poids valent 1. Les poids doivent être non-négatifs. Attribuer un poids de 2 à une observation revient à répéter deux fois la mesure correspondante. Si l'option « libellés des colonnes » est activée, assurez-vous que l'en-tête (première ligne) a également été sélectionné.

Nombre de classes :

De : Entrer un nombre compris entre 1 et 25.

À : Entrer un nombre compris entre 1 et 25.

N.B. : pour spécifier un nombre fixe de classes K, introduire de K à K. Par exemple, pour un modèle à 2 classes : de 2 à 2.

Utiliser des feuilles séparées : Activez cette option si vous souhaitez que le programme fournisse une feuille par modèle de classification estimé. Une feuille supplémentaire contiendra un résumé de statistiques portant sur tous les modèles estimés.

Onglet **Options** :

L'estimation des paramètres se fait grâce à un algorithme itératif qui démarre avec un algorithme **EM** (Expectation-Maximisation), jusqu'à ce que le nombre d'itérations EM ou le critère de convergence EM (**Tolérance(EM)**) sont atteints. Puis, le programme bascule sur des itérations de Newton-Raphson (**NR**) jusqu'à ce que le nombre maximal d'itérations NR ou si le critère global de convergence (**Tolérance**) sont atteints. Le programme peut également arrêter les itérations si la variation du log-posterior est négligeable ($< 10^{-12}$). Une alerte s'affiche si un des éléments du gradient est supérieur à 10^{-3}.

Il peut être plus efficace d'utiliser l'algorithme EM uniquement, dans le cas de modèles impliquant un grand nombre de paramètres. Ceci peut-être indiqué en paramétrant les itérations de Newton-Raphson à 0. Pour les modèles volumineux, il peut être utile de supprimer le calcul d'erreurs standard (et des statistiques de Wald associées) dans l'onglet Sorties.

Convergence

Tolérance(EM) : La tolérance EM est la somme des valeurs absolues de changements relatifs de valeurs de paramètres au cours d'une itération EM. Elle détermine le moment où le

programme bascule d'itérations EM à des itérations NR (si le nombre d'itérations NR est paramétré à > 0). L'augmentation de la tolérance EM impliquera un changement plus rapide de EM à NR. Valeurs acceptées : réels positifs. Valeur par défaut : 0.01. Des valeurs comprises entre 0.01 et 0.1 (1% et 10%) sont raisonnables.

Tolérance : La tolérance globale est la somme des valeurs absolues de changements relatifs de valeurs de paramètre au cours d'une itération. Elle détermine le moment où le programme doit arrêter les itérations. Valeurs acceptées : réels positifs. Valeur par défaut : $1.0x10-8$, ce qui correspond à un critère de convergence assez sévère.

N.B. : Si des itérations uniquement EM sont paramétrées, la tolérance correspond au maximum entre Tolérance(EM) et tolérance globale.

Itérations :

EM : Nombre maximal d'itérations EM. Valeurs acceptées : entiers positifs. Valeur par défaut : 250. Si le modèle ne converge pas au bout de 250 itérations, cette valeur doit être augmentée. Il est également conseillé d'augmenter cette valeur si les itérations de Newton-Raphson sont paramétrées à 0.

Newton-Raphson : Nombre maximal d'itérations Newton-Raphson (NR). Valeurs acceptées : entiers positifs. Valeur par défaut : 50. Une valeur de 0 entraîne l'utilisation exclusive de EM par XLSTAT-LG, ce qui a pour conséquence une convergence plus rapide de modèles contenant un grand nombre de paramètres ou de modèles impliquant des variables continues.

Valeurs initiales

Le meilleur moyen d'éviter d'aboutir à une solution locale est d'utiliser plusieurs jeux de valeurs initiales. L'utilisation de plusieurs jeux est automatisée. Cette procédure augmente considérablement la probabilité de trouver la solution globale, mais ne garantit pas pour autant que cette solution soit trouvée au cours d'un seul essai. Les options qui suivent permettent d'augmenter le nombre de jeux aléatoires et/ou le nombre d'itérations par jeu, afin de diminuer la probabilité d'obtenir des solutions locales.

Jeux aléatoires : nombre de jeux aléatoires de valeurs initiales à utiliser par l'algorithme itératif d'estimation. Une augmentation du nombre de jeux de valeurs initiales de paramètres réduit la probabilité de converger vers une solution locale plutôt que globale. Valeurs acceptées : entiers positifs. Une valeur de 0 ou de 1 entraîne l'utilisation d'un seul jeu de valeurs initiales. Valeur par défaut : 16.

Itérations : choisissez le nombre d'itérations à effectuer pour chaque jeu de valeurs initiales. XLSTAT-LG effectue ce nombre d'itérations pour chaque jeu de valeurs initiales puis deux fois ce nombre pour 10% des meilleurs jeux. Pour certains modèles, un nombre bien supérieur à 20 itérations peut être nécessaire pour éviter les solutions locales.

Graine : La valeur par défaut de 123456789 signifie que la graine est obtenue au cours des estimations via un pseudo-générateur de nombres aléatoires basé sur l'heure. En indiquant un entier négatif différent de 0, le même résultat sera obtenu à chaque fois pour les mêmes données.

Si vous souhaitez introduire une graine particulière, telle que la meilleure graine ne départ obtenue dans la partie résumé des sorties d'un modèle estimé précédemment désactivez les jeux aléatoires (en spécifiant jeux aléatoires = 0).

Tolérance : Il s'agit du critère de convergence du calcul effectué en utilisant les différents jeux de valeurs initiales. La définition de cette tolérance est identique à celle utilisée dans le cadre des itérations EM et Newton-Raphson.

Constantes de Bayes :

Les options de Bayes peuvent être utilisées pour éliminer la possibilité d'obtenir des solutions limites. Valeurs acceptées : réels positifs. Des constantes de Bayes peuvent être introduites pour trois situations différentes :

Latente : Valeur par défaut : 1. Augmentez cette valeur pour augmenter le poids attribué au prior de Dirichlet, utilisé pour éviter l'occurrence de zéros limites dans l'estimation de la distribution latente. Cette valeur peut être interprétée comme un nombre total d'observations ajoutées distribuées équitablement parmi les classes (et les formes de covariables).

Catégoriel : Valeur par défaut : 1. Augmenter cette valeur pour augmenter le poids alloué au prior de Dirichlet utilisé dans l'estimation de modèles multinomiaux impliquant des variables ordinales ou nominales. Cette valeur peut être interprétée comme un nombre total d'observations ajoutées aux cellules dans les modèles pour les indicateurs afin d'éviter l'occurrence de solutions limites.

Variance de l'erreur : Valeur par défaut : 1. Augmentez cette valeur pour augmenter le poids attribué au prior inverse-Wishart utilisé pour estimer la matrice de variance-covariance de l'erreur dans les modèles impliquant des indicateurs continus. Cette valeur peut être interprétée comme un nombre de pseudo-observations rajoutées aux données, chaque pseudo-observation étant associée à un carré d'erreur égal à la variance totale de l'indicateur concerné. Ce prior évite l'occurrence de variances nulles.

Pour les détails techniques, voir la section 7.3 de Vermunt & Magidson (2013a).

Indépendant de la classe

(Co)variance des erreurs : Activez cette option pour indiquer que la covariance des erreurs est forcée à être homogène parmi les classes (= indépendante de la classe considérée). Notez que cette option ne fonctionne que pour les paires d'indicateurs continus pour lesquels des effets directs ont été inclus dans le modèle (voir option **effets directs** dans l'onglet **général**).

Onglet **Données manquantes** :

Ne pas accepter les valeurs manquantes : activez cette option pour que XLSTAT empêche la poursuite des calculs si des valeurs manquantes sont détectées.

Supprimer les observations : activez cette option pour supprimer les observations comportant des données manquantes.

Onglet **Sorties** :

Suite aux calculs, un résumé standard du modèle est généré. L'onglet sortie permet de programmer en plus les sorties suivantes :

Statistiques descriptives : activez cette option pour afficher les statistiques descriptives pour les variables sélectionnées.

Statistiques : Activez cette option pour afficher les statistiques suivantes associées au(x) modèle(s).

Khi² : activez cette option pour afficher diverses statistiques basées sur le khi² et liées à l'ajustement du modèle

Log-vraisemblance : activez cette option si vous souhaitez obtenir les statistiques associées au log de la vraisemblance.

Classification : activez cette option pour afficher les tableaux de classification (tabulations croisées pour les classes modales et probabilistes).

Profil : activez cette option pour afficher les probabilités ou moyennes associées à chaque indicateur.

- La première ligne contient la taille de chaque classe.
- Pour les variables nominales et ordinales, le corps du tableau contient les probabilités conditionnelles marginales reflétant le lien entre les classes et les différentes catégories de chaque variable. La somme de ces probabilités est égale à 1 au sein de chaque combinaison classe/variable (colonnes).
- Pour les variables continues, le corps du tableau contient des moyennes. Les moyennes sont affichées pour les variables type ordinales, en plus des probabilités.

Erreurs standard : activez cette option pour afficher les erreurs standard. La méthode de calcul standard (hessienne) utilise les dérivées du second ordre de la fonction de log-vraisemblance, appelée matrice hessienne.

Résidus bivariés : activez cette option pour afficher les résidus bivariés.

Effectifs/résidus : activez cette option pour afficher les effectifs observés et attendus, ainsi que les résidus standardisés. Cette option n'est pas disponible si une ou plusieurs variables sont continues.

Détails des itérations : activez cette option pour afficher des informations techniques liées à la performance des algorithmes d'estimation (EM et NR). Parmi ces informations : valeurs de log-posterior et de log- vraisemblance au moment de la convergence. Ces détails comprennent également des messages d'avertissement concernant une non-convergence, des paramètres non identifiés, ainsi que des solutions limites.

Equation de scoring : activez cette option pour afficher l'équation de scoring constituée de coefficients de régression associés à un modèle logit multinomial. Les scores sont des

prédictions de logits associés à chaque classe latente t. Par exemple, pour des réponses $Y_1 = j, Y_2 = k, Y_3 = m, Y_4 = s$ à 4 variables nominales, le logit associé à la classe t est :

$$Logit(t) = a[t] + b_1[j,t] + b_2[k,t] + b_3[m,t] + b_4[s,t]$$

Ainsi, pour obtenir les probabilités d'appartenance à la classe t_0, nous utilisons la formule suivante :

$$Prob(classe[t = t_0]|Y_1 = j, Y_2 = k, Y_3 = m, Y_4 = s) = exp\left(\frac{Logit[t_0]}{\sum_t exp(Logit[t]}\right)$$
$$= \frac{exp(a[t_0] + b_1[j,t_0] + b_2[k,t_0] + b_3[m}{\sum_t exp(a[t] + b_1[j,t] + b_2[k,t] + b_3[}$$

Pour plus de détails, merci de consulter le tutoriel.

Onglet **Graphiques** :

Profil des classes : Le profil des classes est construit à partir des probabilités conditionnelles pour les variables nominales et des moyennes pour les autres variables (voir profil dans onglet sortie). Les quantités associées aux classes sélectionnées sont tracées et connectées. Les variables ordinales, continues, comptages et covariables numériques subissent une transformation d'échelle de manière à ce qu'elles soient toujours comprises entre 0 et 1 : pour chaque variable et au sein de chaque classe, la valeur observée minimale est soustraite des moyennes et les résultats sont divisés par l'étendue des valeurs observées (maximum – minimum). L'avantage de cette transformation est de permettre de visualiser ces nombres sur la même échelle que les probabilités associées aux variables nominales. Pour les variables nominales comprenant plus de 2 catégories, toutes les catégories sont affichées simultanément. Pour les variables binaires déclarées en nominales, seule la dernière catégorie est affichée.

Résultats

Feuille résumé

Statistiques descriptives : Dans ce tableau sont affichées les statistiques descriptives correspondant aux différentes variables (indicateurs).

Statistiques pour chaque modèle :

- **Nom du modèle** : Les noms de modèles correspondent aux nombres de classes correspondants.
- **LV** : Log de vraisemblance pour le modèle en cours.
- **BIC(LV), AIC(LV), AIC3(LV)** : BIC, AIC et AIC3 (basés sur LV). En plus de l'ajustement du modèle, ces statistiques prennent en compte sa parcimonie (DDL ou nombre de

paramètres). Dans le cadre de la comparaison de modèles, le meilleur modèle est associé à aux BIC, AIC ou AIC3 les plus faibles.

- **Nombre de paramètres** : Nombre de paramètres.
- **V²** : χ^2 associé au rapport de vraisemblance. Cette statistique est absente si le modèle contient au moins 1 indicateur continu.
- **DDL** : Degrés de liberté associés au V^2.
- **p-value** : p-value associée au V^2.
- **Err.Class.** : Erreur de classification attendue. Proportion de cas mal classés selon le mode (c'est-à-dire assignation des cas aux classes pour lesquelles la probabilité d'appartenance est la plus élevée).

Sorties pour chaque modèle

Statistiques descriptives :

- **Nombre d'observations** : Nombre d'observations utilisées dans l'estimation du modèle. Ce nombre doit être inférieur au nombre d'observations contenues dans les données au cas où les données manquantes ont été exclues.
- **Nombre de paramètres** : Nombre de paramètres distincts estimés.
- **Graine (nombres aléatoires)** : Graine nécessaire à la reproduction de ce modèle.
- **Meilleure graine** : Graine unique permettant de reproduire ce modèle plus rapidement en utilisant un nombre de jeux aléatoires de valeurs initiales = 0.

Résumé de l'estimation :

- **Itérations EM** : Nombre d'itérations EM utilisées.
- **Log-posterior** : Valeur du log-posterior.
- **V²** : Valeur d'ajustement du rapport de vraisemblance.
- **Valeur de convergence finale** : Valeur de convergence finale.
- **Itérations Newton-Raphson** : Nombre d'itérations Newton-Raphson utilisées.
- **Log-posterior** : Valeur du log-posterior.
- **V²** : Valeur d'ajustement du rapport de vraisemblance.
- **Valeur de convergence finale** : Valeur de convergence finale.

Statistiques pour le Khi² :

- **DDL** : Degrés de liberté associés au modèle.
- **V²** : Valeur d'ajustement du rapport de vraisemblance. Si elle est paramétrée, la p-value bootstrap pour le V² est affichée.
- **X² et Cressie-Read** : Alternatives au V^2, qui devraient fournir une p-value similaire d'après la théorie des grands échantillons, si le modèle spécifié est valide et que les données ne sont pas rares.
- **BIC(LV), AIC(LV), AIC3(LV)** : BIC, AIC et AIC3 (basés sur LV). En plus de l'ajustement du modèle, ces statistiques prennent en compte sa parcimonie (DDL ou nombre de paramètres). Dans le cadre de la comparaison de modèles, le meilleur modèle est associé à aux BIC, AIC ou AIC3 les plus faibles.
- **SABIC (LV)** : BIC ajusté selon la taille de l'échantillon. Ce critère se calcule de manière similaire, mais en remplaçant $log(N)$ par $log(\frac{(N+2)}{24})$.
- **Indice de dissimilarité** : Mesure reflétant la distance entre les fréquences de cellules observées et estimées. Elle indique la proportion d'échantillon à déplacer d'une cellule à une autre afin d'obtenir un ajustement parfait.

Statistiques de Log-vraisemblance

- **Log-vraisemblance(LV)** : Logarithme népérien de la vraisemblance.
- **Log-prior** : terme issu de la fonction maximisée dans l'estimation de paramètres associée aux constantes de Bayes. Ce terme est égal à 0 si toutes les constantes de Bayes sont = 0.
- **Log-posterior** : terme issu de la fonction maximisée dans l'estimation de paramètres. La valeur du log-posterior est la somme du log-vraisemblance et du log-prior.
- **BIC, AIC, AIC3 et CAIC (basés sur le LV)** : Ces statistiques (critères d'information) trouvent un compromis entre ajustement et parcimonie en corrigeant le LV en fonction du nombre de paramètres présents dans le modèle. Dans le cadre de la comparaison de modèles, le meilleur modèle est associé à aux BIC, AIC ou AIC3 les plus faibles.
- **SABIC (LV)** : BIC ajusté selon la taille de l'échantillon. Ce critère se calcule de manière similaire, mais en remplaçant $log(N)$ par $log(\frac{(N+2)}{24})$.

Statistiques de classification :

- **Erreurs de classification** : Lorsque la classification des cas est basée sur le mode (c'est-à-dire assignation des cas aux classes pour lesquelles la probabilité d'appartenance est la plus élevée), cette statistique décrit la proportion de cas estimés en tant que mal classés. Plus cette valeur est proche de zéro, meilleur est le modèle.
- **Réduction des erreurs (Lambda), R² d'entropie, R² standard** : Ces pseudo-R^2 reflètent la qualité de prédiction des appartenances à des classes selon les variables

observées. Plus ces statistiques se rapprochent de 1, meilleure est la qualité prédictive du modèle.

- **Log-vraisemblance de la classification** : valeur de log-vraisemblance sous l'hypothèse que l'appartenance réelle aux classes est connue.
- **AWE** : Similaire au BIC, mais prend aussi en compte plus la performance de classification.
- **EN** : Entropie.
- **CLC** : CL*2
- **ICL_BIC** : BIC-2*EN

Tableau de classification :

- **Modale** : Tableau croisé des affectations à des classes selon le mode.
- **Proportionnelle** : Tableau croisé des affectations à des classes selon la probabilité d'appartenance.

Profil :

- **Effectif de classe** : taille de chaque classe.
- **Modalités** : Le corps du tableau contient les probabilités conditionnelles (marginales) indiquant la manière dont les classes sont liées aux variables indicatrices nominales ou ordinales. La somme de ces probabilités est 1. Pour les variables indicatrices continues, le corps du tableau contient les moyennes, mais pas les probabilités. Pour les variables ordinales, les moyennes sont indiquées en plus des probabilités.
- **Erreurs standard** : Erreurs standard associées aux probabilités conditionnelles (marginales).
- **Profil des classes** (graphique) : Représentation graphique des probabilités et moyennes inclus dans le tableau de profil.

Effectifs Résidus :

- Tableau de fréquences (et résidus) observés / estimés. N.B. : les résidus dont l'amplitude dépasse 2 sont statistiquement significatifs. Cette sortie n'apparait pas au cas où le modèle inclut au moins un indicateur continu.

Résidus bivariés :

- **Variables** : Tableau contenant les résidus bivariés. Des valeurs élevées de résidus bivariés suggèrent une violation de l'hypothèse d'indépendance locale.

Equation de scoring : coefficients de régression associés au modèle logit multinomial.

Classification : Affiche pour chaque observation l'appartenance a posteriori aux classes ainsi que les affectations modales, selon le modèle.

Messages d'alerte :

MESSAGE : nombre négatif de degrés de liberté.

Cette alerte indique que le modèle contient plus de paramètres que de cellules. Une condition nécessaire (mais pas suffisante) pour l'identification des paramètres d'un modèle sur classes latentes est que le nombre de degrés de liberté soit positif. Cette alerte montre donc que le modèle n'est pas identifié. Utiliser un modèle avec moins de classes latentes.

MESSAGE : # paramètre(s) limite(s) ou non-identifié(s)

Cette alerte est dérivée du rang de la matrice d'information (hessienne ou son approximation par le produit extérieur). La présence de paramètres non- identifiés engendre une matrice d'information qui ne sera pas de plein rang. Le nombre affiché est la déficience de rang, indiquant le nombre de paramètres non-identifiés.

Notez que deux problèmes sont associés à la vérification de l'identification. Le premier est que les estimations limites engendrent elles aussi des déficiences de rang. En d'autres termes, nous ne pouvons pas savoir si une déficience de rang est causée par des limites ou par des paramètres non- identifiés. Les constantes de Bayes empêchent les limites d'apparaître, ce qui résout le premier problème lié à ce message. Cependant, un second problème provient du fait que cette vérification d'identification ne peut pas toujours détecter la non-identification lorsque des constantes de Bayes sont utilisées. En d'autres termes, des constantes de Bayes peuvent faire apparaître en tant qu'identifiés des modèles non-identifiés en réalité.

MESSAGE : nombre maximum d'itérations atteint sans convergence

Cette alerte apparaît lorsque le nombre maximal d'itérations EM et Newton- Raphson est atteint avant de vérifier le critère de tolérance. Si le critère de tolérance (très strict par défaut) est presque atteint, la solution est probablement fiable. Sinon, il est recommandé de ré-estimer le modèle avec une tolérance EM plus sévère et/ou un nombre d'itérations EM plus élevé. Cela retarde le basculement entre itérations EM et itérations de Newton-Raphson. Le nombre d'itérations de Newton-Raphson par défaut (50) est en général suffisant.

MESSAGE : la procédure d'estimation n'a pas convergé (# gradients supérieur à $1.0e-3$)

Ce message peut être lié au message précédent, auquel cas la même solution peut être envisagée. L'absence d'affichage du message précédent suggère un problème plus grave de non-convergence. Les algorithmes pourraient avoir été piégés dans une région très aplatie de l'espace des paramètres (point selle). La meilleure solution est de ré-estimer le modèle avec une graine différente, et si possible avec un plus grand nombre de jeux de valeurs initiales et d'itérations par jeu.

Exemple

Un exemple de classification par les classes latentes est disponible sur le Centre d'aide XLSTAT :

http://www.xlstat.com/demo-lccf.htm

Bibliographie

Vermunt J.K. (2010). Latent class modeling with covariates: Two improved three-step approaches. Political Analysis, 18, 450-469. Link: http://members.home.nl/jeroenvermunt/lca_three_step.pdf

Vermunt J.K. and Magidson, J. (2005). Latent GOLD 4.0 User's Guide. Belmont, MA: Statistical Innovations Inc. http://www.statisticalinnovations.com/technicalsupport/LGusersguide.pdf

Vermunt J.K. and Magidson, J. (2013a). Technical Guide for Latent GOLD 5.0: Basic, Advanced, and Syntax. Belmont, MA: Statistical Innovations Inc. http://www.statisticalinnovations.com/technicalsupport/LGtechnical.pdf

Vermunt J.K. and Magidson J. (2013b). Latent GOLD 5.0 Upgrade Manual. Belmont, MA: Statistical Innovations Inc.

http://statisticalinnovations.com/technicalsupport/LG5manual.pdf

Régression par les classes latentes

Cet outil permet de classifier des cas (observations, individus) au sein de groupes (classes latentes) différents selon un ou plusieurs paramètres, dans le cadre de modèles de classification par les classes latentes (CL). Les modèles de régression CL classifient les cas et fournissent en plus des estimations de coefficients de régression basés sur des modèles linéaires, logistiques, multinomiaux, de comptages binomiaux ou de Poisson.

Dans cette section :

Description

XLSTAT-LG est un outil de classification puissant basé sur deux modules de Latent GOLD® 5.0 : **modèles de classification par les classes latentes** et **modèles de régression sur classes latentes**.

Les analyses en classes latentes (ACL) impliquent la construction de classes latentes (CL), qui sont des sous-groupes ou segments non-observés (latents) de cas (observations, individus). Les CL sont construites en se basant sur les réponses observées (manifestes) des cas sur un ensemble de variables indicatrices. Les cas se trouvant dans la même classe latente sont similaires sur le plan de leurs réponses tandis que les cas se trouvant dans des classes différentes le sont moins. Formellement, les classes sont représentées par K catégories distinctes d'une variable nominale latente X. Comme X est catégorielle, la modélisation LC se distingue d'approches plus traditionnelles telles que l'analyse factorielle, les modèles d'équations structurelles, ainsi que les modèles de régression impliquant des effets aléatoires. Ces approches sont plutôt basées sur des variables latentes continues.

XLSTAT-LG comprend deux modules pour l'estimation de deux structures de modèles : **modèles de classification CL** et **modèles de régression CL**. Ces deux modules sont utilisables dans des domaines plus ou moins différents. Dans la suite de l'aide, le terme « segments » se réfère à « classes latentes ».

Le modèle de régression CL :

- Est utilisé pour prédire une variable dépendante (à expliquer) en fonction de variables prédictives (explicatives).

- Met en jeu une variable latente X à K catégories (modèle CL)
- Chaque catégorie représente une sous-population (segment) homogène associée à des coefficients de régression identiques.
- Chaque cas (individu) peut être associé à plusieurs mesures (régression CL avec mesures répétées).
- Le type de modèle est estimé en fonction du type de variable dépendante :
- Variable continue : régression linéaire (avec résidus distribués normalement).
- Variable nominale avec plus de deux modalités : régression logistique multinomiale.
- Variable ordinale avec plus de deux niveaux ordonnés : régression logistique ordinale basée sur des catégories adjacentes.
- Variable type comptages : régression log-linéaire de Poisson.
- Variable binomiale : régression logistique binomiale.

Des variables dépendantes dichotomiques peuvent être analysées pour les types nominal, ordinal ou binomial. Cela ne produit aucune différence au niveau des résultats.

Quel que soit le type de modèle :

- Des statistiques de diagnostic sont disponibles pour aider à déterminer le nombre de classes latentes.
- Des variables peuvent être incluses dans les modèles contenant $K > 1$ classes, afin d'améliorer la classification de chaque cas au sein des segments les plus vraisemblables.

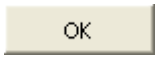
Parmi les avantages de la méthode par rapport à des approches traditionnelles de classification, citons la présence de critères de sélection de modèles et des classifications probabilistes. Les probabilités d'appartenance a posteriori sont estimées directement à partir des paramètres du modèle et sont utilisées pour assigner chaque cas à la classe modale correspondante, à savoir la classe associée à la probabilité d'appartenance a posteriori la plus élevée.

Equation de scoring : ces modèles fournissent une équation de scoring permettant de calculer les probabilités d'appartenance a posteriori directement à partir de variables observées (indicateurs). Cette équation peut être utilisée pour affilier de nouveaux cas à la classe la plus vraisemblable. Cette fonctionnalité est exclusive des modèles de classification CL.

L'équation de scoring est obtenue sous forme de cas spécial de la troisième étape dans la construction des modèles de classification CL (Vermunt 2010). La première étape implique une estimation des paramètres du modèle. Au cours de la deuxième étape, les cas sont assignés à des classes selon leurs probabilités d'appartenance a posteriori. Dans le cadre de la troisième étape, les classes latentes sont utilisées en tant que prédicteurs ou variables dépendantes pour des analyses ultérieures. Pour plus de détails, cf. Vermunt et Magidson (2013).

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.

Onglet **Général** :

Y / variables dépendantes : sélectionnez la variable dépendante. Si le libellé des variables a été sélectionné, veuillez vérifier que l'option « Libellés des colonnes » est activée.

N.B. Si plusieurs variables dépendantes sont sélectionnées, une analyse de régression sera faite par variable dépendante. Des sorties séparées seront produites. Les variables dépendantes doivent être de même type.

Type de réponse : sélectionner le type de variable dépendante : nominale, ordinale, continue, binomiale ou comptages.

- **Nominale.** Cette option doit être utilisée pour une variable catégorielle dont les catégories ne peuvent pas être ordonnées. Ce choix entraîne l'utilisation d'un modèle logit multinomial.

- **Ordinale.** Cette option doit être utilisée pour une variable catégorielle dont les catégories sont ordonnées. Ce choix entraîne l'utilisation d'un modèle de régression logistique ordinale basée sur des catégories adjacentes.

- **Continue.** Cette option doit être utilisée pour une variable quantitative continue. Ceci entraîne l'utilisation d'un modèle de régression linéaire normal.

- **Binomiale.** Cette option doit être utilisée pour une variable représentant des comptages binomiaux. Ce choix entraîne l'utilisation d'un modèle logistique binomial. Vous pouvez inclure une variable d'exposition (cf. exposition, paragraphe qui suit). Durant la lecture des données, XLSTAT-LG vérifie que la variable d'exposition est supérieure aux comptages observés.

- **Comptages.** Cette option doit être utilisée pour une variable représentant des comptages de Poisson. Ce choix entraîne l'utilisation d'un modèle de Poisson. Là encore, il est possible d'inclure une variable d'exposition (cf. paragraphe suivant).

Exposition. Ce champ apparaît lorsque le type de réponse sélectionné est Binomiale ou Comptages.

L'exposition est spécifiée par une variable sélectionnée parmi les données ou par une valeur unique. L'utilisation d'une variable permet de faire varier l'exposition à travers les cas (individus). Valeur par défaut : 1. Il s'agit d'une valeur souvent utilisée pour représenter une exposition pour les comptages de Poisson.

Pour une réponse type binomiale, la valeur de la variable dépendante représente un nombre de « succès » parmi N essais. Dans ce cas, l'exposition représente le nombre d'essais (N), et par conséquent ne doit jamais prendre une valeur inférieure à la valeur correspondante au sein de la variable dépendante (pas plus de succès que d'essais). Par ailleurs, elle doit être supérieure à la valeur constante par défaut (1). Durant la lecture des données, XLSTAT-LG vérifie chaque observation et renvoie un message si ces conditions ne sont pas vérifiées pour toutes les données.

Une variable d'exposition (plutôt qu'une constante) doit être sélectionnée si le nombre d'essais varie d'un cas à un autre.

Variables explicatives. Sélectionnez les variables à utiliser en tant que prédicteurs de la variable dépendante. Deux types sont possibles : variables nominales ou numériques. Si aucun prédicteur n'est sélectionné, le modèle ne contiendra que l'estimation d'une constante (intercept).

- **Numériques.** Ce champ doit contenir les prédicteurs et covariables ordinales ou continues.

- **Nominales.** Ce champ doit contenir des variables catégorielles dont les catégories ne peuvent pas être ordonnées.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des colonnes : activez cette option si la première ligne des données sélectionnées (variables dépendantes et explicatives) contient des libellés.

Libellés des observations : Activez cette option si des libellés sont disponibles pour les N observations. Puis sélectionnez les données correspondantes. Si l'option « libellés des colonnes » a été activée, la première cellule de la sélection doit comprendre un en-tête.

Pour les mesures répétées (plusieurs mesures par cas), les libellés des observations servent d'identifiants qui regrouperont ensemble les mesures pour chaque cas. Si vous n'activez pas cette option, des libellés seront automatiquement créés (Obs1, Obs2, …) et il y aura autant de cas que d'observations.

Poids des observations : Activez cette option is vous souhaitez attribuer un poids aux observations. Si vous n'activez pas cette option, tous les poids valent 1. Les poids doivent être non-négatifs. Attribuer un poids de 2 à une observation revient à répéter deux fois la mesure correspondante. Si l'option « libellés des colonnes » est activée, assurez-vous que l'en-tête (première ligne) a également été sélectionné.

Nombre de classes :

De : Entrer un nombre compris entre 1 et 25.

À : Entrer un nombre compris entre 1 et 25.

N.B. : pour spécifier un nombre fixe de classes K, introduire de K à K. Par exemple, pour un modèle à 2 classes : de 2 à 2.

Utiliser des feuilles séparées : Activez cette option si vous souhaitez que le programme fournisse une feuille par modèle de classification estimé. Une feuille supplémentaire contiendra un résumé de statistiques portant sur tous les modèles estimés.

Onglet **Options** :

L'estimation des paramètres se fait grâce à un algorithme itératif qui démarre avec un algorithme **EM** (Expectation-Maximisation), jusqu'à ce que le nombre d'itérations EM ou le critère de convergence EM (**Tolérance(EM)**) sont atteints. Puis, le programme bascule sur des itérations de Newton-Raphson (**NR**) jusqu'à ce que le nombre maximal d'itérations NR ou si le critère global de convergence (**Tolérance**) sont atteints. Le programme peut également arrêter

les itérations si la variation du log-posterior est négligeable ($< 10^{-12}$). Une alerte s'affiche si un des éléments du gradient est supérieur à 10^{-3}.

Il peut être plus efficace d'utiliser l'algorithme EM uniquement, dans le cas de modèles impliquant un grand nombre de paramètres. Ceci peut-être indiqué en paramétrant les itérations de Newton-Raphson à 0. Pour les modèles volumineux, il peut être utile de supprimer le calcul d'erreurs standard (et des statistiques de Wald associées) dans l'onglet Sorties.

Convergence

Tolérance(EM): La tolérance EM est la somme des valeurs absolues de changements relatifs de valeurs de paramètres au cours d'une itération EM. Elle détermine le moment où le programme bascule d'itérations EM à des itérations NR (si le nombre d'itérations NR est paramétré à > 0). L'augmentation de la tolérance EM impliquera un changement plus rapide de EM à NR. Valeurs acceptées : réels positifs. Valeur par défaut : 0.01. Des valeurs comprises entre 0.01 et 0.1 (1% et 10%) sont raisonnables.

Tolérance: La tolérance globale est la somme des valeurs absolues de changements relatifs de valeurs de paramètre au cours d'une itération. Elle détermine le moment où le programme doit arrêter les itérations. Valeurs acceptées : réels positifs. Valeur par défaut : $1.0x10^{-8}$, ce qui correspond à un critère de convergence assez sévère.

Itérations :

EM : Nombre maximal d'itérations EM. Valeurs acceptées : entiers positifs. Valeur par défaut : 250. Si le modèle ne converge pas au bout de 250 itérations, cette valeur doit être augmentée. Il est également conseillé d'augmenter cette valeur si les itérations de Newton-Raphson sont paramétrées à 0.

Newton-Raphson : Nombre maximal d'itérations Newton-Raphson (NR). Valeurs acceptées : entiers positifs. Valeur par défaut : 50. Une valeur de 0 entraîne l'utilisation exclusive de EM par XLSTAT-LG, ce qui a pour conséquence une convergence plus rapide de modèles contenant un grand nombre de paramètres ou de modèles impliquant des variables continues.

Valeurs initiales

Le meilleur moyen d'éviter d'aboutir à une solution locale est d'utiliser plusieurs jeux de valeurs initiales. L'utilisation de plusieurs jeux est automatisée. Cette procédure augmente considérablement la probabilité de trouver la solution globale, mais ne garantit pas pour autant que cette solution soit trouvée au cours d'un seul essai. Les options qui suivent permettent d'augmenter le nombre de jeux aléatoires et/ou le nombre d'itérations par jeu, afin de diminuer la probabilité d'obtenir des solutions locales.

Jeux aléatoires : nombre de jeux aléatoires de valeurs initiales à utiliser par l'algorithme itératif d'estimation. Une augmentation du nombre de jeux de valeurs initiales de paramètres réduit la probabilité de converger vers une solution locale plutôt que globale. Valeurs acceptées : entiers positifs. Une valeur de 0 ou de 1 entraîne l'utilisation d'un seul jeu de valeurs initiales. Valeur par défaut : 16.

Itérations : choisissez le nombre d'itérations à effectuer pour chaque jeu de valeurs initiales. XLSTAT-LG effectue ce nombre d'itérations pour chaque jeu de valeurs initiales puis deux fois

ce nombre pour 10% des meilleurs jeux. Pour certains modèles, un nombre bien supérieur à 20 itérations peut être nécessaire pour éviter les solutions locales.

Graine : La valeur par défaut de 123456789 signifie que la graine est obtenue au cours des estimations via un pseudo-générateur de nombres aléatoires basé sur l'heure. En indiquant un entier négatif différent de 0, le même résultat sera obtenu à chaque fois pour les mêmes données.

Si vous souhaitez introduire une graine particulière, telle que la meilleure graine ne départ obtenue dans la partie résumé des sorties d'un modèle estimé précédemment désactivez les jeux aléatoires (en spécifiant jeux aléatoires = 0).

Tolérance : Il s'agit du critère de convergence du calcul effectué en utilisant les différents jeux de valeurs initiales. La définition de cette tolérance est identique à celle utilisée dans le cadre des itérations EM et Newton-Raphson.

Constantes de Bayes :

Les options de Bayes peuvent être utilisées pour éliminer la possibilité d'obtenir des solutions limites. Valeurs acceptées : réels positifs. Des constantes de Bayes peuvent être introduites pour trois situations différentes :

Latente : Valeur par défaut : 1. Augmentez cette valeur pour augmenter le poids attribué au prior de Dirichlet, utilisé pour éviter l'occurrence de zéros limites dans l'estimation de la distribution latente. Cette valeur peut être interprétée comme un nombre total d'observations ajoutées distribuées équitablement parmi les classes (et les formes de covariables).

Catégoriel : Valeur par défaut : 1. Augmenter cette valeur pour augmenter le poids alloué au prior de Dirichlet utilisé dans l'estimation de modèles multinomiaux impliquant des variables ordinales ou nominales. Cette valeur peut être interprétée comme un nombre total d'observations ajoutées aux cellules dans les modèles pour les indicateurs afin d'éviter l'occurrence de solutions limites.

Comptages de Poisson : Valeur par défaut : 1. Ce prior équivaut à l'addition d'un nombre spécifique d'évènements aux données sans changer le taux global de Poisson. En d'autres termes, le nombre d'expositions est ajusté en fonction. Ce prior évite l'occurrence de solutions limites pour les modèles impliquant des variables dépendantes type comptages de Poisson.

Variance de l'erreur : Valeur par défaut : 1. Augmentez cette valeur pour augmenter le poids attribué au prior inverse-Wishart utilisé pour estimer la matrice de variance-covariance de l'erreur dans les modèles impliquant des indicateurs continus. Cette valeur peut être interprétée comme un nombre de pseudo-observations rajoutées aux données, chaque pseudo-observation étant associée à un carré d'erreur égal à la variance totale de l'indicateur concerné. Ce prior évite l'occurrence de variances nulles.

Pour les détails techniques, voir la section 7.3 de Vermunt & Magidson (2013a).

Indépendant de la classe

Différentes restrictions sont disponibles pour les effets des constantes et des prédicteurs. Pour les modèles dont la variable dépendante est continue, il est par ailleurs possible d'établir des restrictions pour les variances des erreurs.

- **Variances des erreurs :** Cette option force les covariances des erreurs à être égales parmi les classes.
- **Prédicteurs (1 ou plus).** Cette option force les prédicteurs à être égaux parmi les classes.

Constante. Cette option force les constantes à être égales parmi les classes.

Onglet **Données manquantes** :

Ne pas accepter les valeurs manquantes : activez cette option pour que XLSTAT empêche la poursuite des calculs si des valeurs manquantes sont détectées.

Supprimer les observations : activez cette option pour supprimer les observations comportant des données manquantes.

Onglet **Sorties** :

Suite aux calculs, un résumé standard du modèle est généré. L'onglet sortie permet de programmer en plus les sorties suivantes :

Statistiques descriptives : activez cette option pour afficher les statistiques descriptives pour les variables sélectionnées.

Statistiques : Activez cette option pour afficher les statistiques suivantes associées au(x) modèle(s).

Khi² : activez cette option pour afficher diverses statistiques basées sur le χ^2 et liées à l'ajustement du modèle

Log-vraisemblance : activez cette option si vous souhaitez obtenir les statistiques associées au log de la vraisemblance.

Classification : activez cette option pour afficher les tableaux de classification (tabulations croisées pour les classes modales et probabilistes).

Paramètres :

Erreurs standard : Activez cette option pour afficher les erreurs standard des paramètres. La méthode de calcul standard (hessienne) utilise les dérivées du second ordre de la fonction de log-vraisemblance, appelée matrice hessienne.

Tests de Wald : Activez cette option pour afficher les statistiques de Wald.

Effectifs/résidus : activez cette option pour afficher les effectifs observés et attendus, ainsi que les résidus standardisés. Cette option n'est pas disponible si une ou plusieurs variables sont

continues.

Détails des itérations : activez cette option pour afficher des informations techniques liées à la performance des algorithmes d'estimation (EM et NR). Parmi ces informations : valeurs de log-posterior et de log- vraisemblance au moment de la convergence. Ces détails comprennent également des messages d'avertissement concernant une non-convergence, des paramètres non identifiés, ainsi que des solutions limites.

Valeurs estimées : Activez cette option pour afficher les informations sur les valeurs prédites (probabilité de réponse à chaque catégorie). Les variables suivantes seront affichées :

- pred_1 – probabilité prédite de réponse à la première catégorie
- pred_2 – probabilité prédite de réponse à la deuxième catégorie
- pred_dep – valeur prédite (moyenne pondérée des scores de catégories, les poids étant les probabilités prédites).

Classification : Activez cette option afin d'afficher un tableau contenant les probabilités d'appartenance a posteriori et l'affectation modale de chaque cas.

Codage nominal:

- **Codage** (option par défaut). Les sorties portant sur les paramètres contiennent un codage d'effet pour les indicateurs nominaux, la variable dépendante, les covariables actives, ainsi que les classes latentes.
- **a1=0.** Activez cette option pour considérer la première catégorie en tant que catégorie de référence (0).
- **an=0.** Activez cette option pour considérer la dernière catégorie en tant que catégorie de référence (0).

Onglet **Graphiques** :

Profil des classes : Activez cette option pour afficher le graphique de profil des classes.

Résultats

Feuille résumé

Statistiques descriptives : Dans ce tableau sont affichées les statistiques descriptives correspondant aux différentes variables (indicateurs).

Statistiques pour chaque modèle :

- **Nom du modèle :** Les noms de modèles correspondent aux nombres de classes correspondants.
- **LV :** Log de vraisemblance pour le modèle en cours.
- **BIC(LV), AIC(LV), AIC3(LV) :** BIC, AIC et AIC3 (basés sur LV). En plus de l'ajustement du modèle, ces statistiques prennent en compte sa parcimonie (DDL ou nombre de paramètres). Dans le cadre de la comparaison de modèles, le meilleur modèle est associé à aux BIC, AIC ou AIC3 les plus faibles.
- **Nombre de paramètres :** Nombre de paramètres.
- **V² :** $\chi - 2$ associé au rapport de vraisemblance. Cette statistique est absente si le modèle contient au moins 1 indicateur continu.
- **DDL :** Degrés de liberté associés au V^2.
- **p-value :** p-value associée au V^2.
- **Err.Class. :** Erreur de classification attendue. Proportion de cas mal classés selon le mode (c'est-à-dire assignation des cas aux classes pour lesquelles la probabilité d'appartenance est la plus élevée).

Sorties pour chaque modèle

Statistiques descriptives :

- **Nombre d'observations :** Nombre d'observations utilisées dans l'estimation du modèle. Ce nombre doit être inférieur au nombre d'observations contenues dans les données au cas où les données manquantes ont été exclues.
- **Nombre de paramètres :** Nombre de paramètres distincts estimés.
- **Graine (nombres aléatoires) :** Graine nécessaire à la reproduction de ce modèle.
- **Meilleure graine :** Graine unique permettant de reproduire ce modèle plus rapidement en utilisant un nombre de jeux aléatoires de valeurs initiales = 0.

Résumé de l'estimation :

- **Itérations EM :** Nombre d'itérations EM utilisées.
- **Log-posterior :** Valeur du log-posterior.
- **V² :** Valeur d'ajustement du rapport de vraisemblance.
- **Valeur de convergence finale** : Valeur de convergence finale.
- **Itérations Newton-Raphson :** Nombre d'itérations Newton-Raphson utilisées.

- **Log-posterior :** Valeur du log-posterior.
- **V² :** Valeur d'ajustement du rapport de vraisemblance.
- **Valeur de convergence finale** : Valeur de convergence finale.

Statistiques pour le Khi² :

- **DDL :** Degrés de liberté associés au modèle.
- **V² :** Valeur d'ajustement du rapport de vraisemblance. Si elle est paramétrée, la p-value bootstrap pour le V^2 est affichée.
- **X² et Cressie-Read :** Alternatives au V^2, qui devraient fournir une p-value similaire d'après la théorie des grands échantillons, si le modèle spécifié est valide et que les données ne sont pas rares.
- **BIC(LV), AIC(LV), AIC3(LV):** BIC, AIC et AIC3 (basés sur LV). En plus de l'ajustement du modèle, ces statistiques prennent en compte sa parcimonie (DDL ou nombre de paramètres). Dans le cadre de la comparaison de modèles, le meilleur modèle est associé à aux BIC, AIC ou AIC3 les plus faibles.
- **SABIC (LV) :** BIC ajusté selon la taille de l'échantillon. Ce critère se calcule de manière similaire, mais en remplaçant $log(N)$ par $log(\frac{(N+2)}{24})$.
- **Indice de dissimilarité :** Mesure reflétant la distance entre les fréquences de cellules observées et estimées. Elle indique la proportion d'échantillon à déplacer d'une cellule à une autre afin d'obtenir un ajustement parfait.

Statistiques de Log-vraisemblance

- **Log-vraisemblance(LV) :** Logarithme népérien de la vraisemblance.
- **Log-prior :** terme issu de la fonction maximisée dans l'estimation de paramètres associée aux constantes de Bayes. Ce terme est égal à 0 si toutes les constantes de Bayes sont = 0.
- **Log-posterior :** terme issu de la fonction maximisée dans l'estimation de paramètres. La valeur du log-posterior est la somme du log-vraisemblance et du log-prior.
- **BIC, AIC, AIC3 et CAIC (basés sur le LV) :** Ces statistiques (critères d'information) trouvent un compromis entre ajustement et parcimonie en corrigeant le LV en fonction du nombre de paramètres présents dans le modèle. Dans le cadre de la comparaison de modèles, le meilleur modèle est associé à aux BIC, AIC ou AIC3 les plus faibles.
- **SABIC (LV) :** BIC ajusté selon la taille de l'échantillon. Ce critère se calcule de manière similaire, mais en remplaçant $log(N)$ par $log(\frac{(N+2)}{24})$.

Statistiques de classification :

- **Erreurs de classification :** Lorsque la classification des cas est basée sur le mode (c'est-à-dire assignation des cas aux classes pour lesquelles la probabilité d'appartenance est la plus élevée), cette statistique décrit la proportion de cas estimés en tant que mal classés. Plus cette valeur est proche de zéro, meilleur est le modèle.
- **Réduction des erreurs (Lambda), R² d'entropie, R² standard :** Ces pseudo-R^2 reflètent la qualité de prédiction des appartenances à des classes selon les variables observées. Plus ces statistiques se rapprochent de 1, meilleure est la qualité prédictive du modèle.
- **Log-vraisemblance de la classification :** valeur de log-vraisemblance sous l'hypothèse que l'appartenance réelle aux classes est connue.
- **AWE :** Similaire au BIC, mais prend aussi en compte plus la performance de classification.
- **EN:** Entropie.
- **CLC:** CL*2
- **ICL_BIC** : BIC-2*EN

Tableau de classification :

- **Modale :** Tableau croisé des affectations à des classes selon le mode.
- **Proportionnelle :** Tableau croisé des affectations à des classes selon la probabilité d'appartenance.

Statistiques de prédiction :

Les colonnes de ce tableau correspondent à :

Base : Erreur de prédiction du modèle de base (aussi appelé modèle nul).

Modèle : Erreur de prédiction du modèle estimé.

R² : Réduction proportionnelle des erreurs dans le modèle estimé en comparaison au modèle de base.

Les lignes de ce tableau correspondent à :

Erreur quadratique : Erreur de prédiction moyenne basée selon le carré de l'erreur.

(Moins) Log-vraisemblance : Erreur de prédiction moyenne selon –log(vraisemblance).

Erreur absolue : Erreur de prédiction moyenne selon l'erreur absolue.

Erreur de prédiction : Erreur de prédiction selon la proportion d'erreurs de prédiction (uniquement pour les variables catégorielles).

Pour plus d'informations techniques, cf. la section 8.1.5 de Vermunt & Magidson (2013a).

Tableau de prédiction : Pour les variables dépendantes nominales ou ordinales, tableau croisant les valeurs observées et les valeurs estimées.

Paramètres :

- **R² :** R^2 spécifiques aux classes et R^2 global. Le R^2 global reflète la qualité de prédiction globale de la variable dépendante par le modèle (même chiffre que dans les statistiques de prédiction). Pour les variables dépendantes ordinales, continues, binomiales et type comptages, il s'agit de R^2 standards. Pour les variables dépendantes nominales, ces R^2 peuvent être assimilés à des moyennes pondérées de R² distincts pour chaque catégorie, chaque catégorie étant traitée comme une variable-réponse dichotomique distincte (1 pour la catégorie et 0 pour le reste).
- **Constante :** Constante de l'équation de régression linéaire.
- **e.s. :** Erreurs standard des paramètres.
- **z-value :** Statistique de test z correspondant aux tests des paramètres.
- **Wald :** Les statistiques de Wald servent à mesurer la significativité statistique d'un ensemble d'estimations de paramètres associées à une variable donnée. Spécifiquement, pour chaque variable, la statistique de Wald teste les hypothèses que chaque estimation de paramètre dans cet ensemble est égale à zéro (pour les variables nominales, l'ensemble inclut un paramètre par catégorie). Pour les modèles de régression, deux statistiques de Wald (**Wald**, **Wald(=)**) sont fournies dans le tableau lorsqu'au moins une classe a été estimée. Pour chaque ensemble d'estimations de paramètres, la statistique Wald(=) prend en compte un sous-ensemble associé à chaque classe et teste les hypothèses que chaque paramètre dans ce sous-ensemble est égal aux paramètres correspondants dans les sous-ensembles associés à chacune des autres classes. En d'autres termes, la statistique Wald(=) teste l'égalité de chaque sous-ensemble d'effets de régression à travers les classes.
- **p-value :** Mesure la significativité des estimations.
- **Moyenne :** Moyennes des coefficients de régression.
- **Ecart-types :** Ecarts-types des coefficients de régression.

Classification : Affiche pour chaque observation l'appartenance a posteriori aux classes ainsi que les affectations modales, selon le modèle.

Messages d'alerte :

MESSAGE : nombre négatif de degrés de liberté.

Cette alerte indique que le modèle contient plus de paramètres que de cellules. Une condition nécessaire (mais pas suffisante) pour l'identification des paramètres d'un modèle sur classes latentes est que le nombre de degrés de liberté soit positif. Cette alerte montre donc que le modèle n'est pas identifié. Utiliser un modèle avec moins de classes latentes.

MESSAGE : # paramètre(s) limite(s) ou non-identifié(s)

Cette alerte est dérivée du rang de la matrice d'information (hessienne ou son approximation par le produit extérieur). La présence de paramètres non- identifiés engendre une matrice d'information qui ne sera pas de plein rang. Le nombre affiché est la déficience de rang, indiquant le nombre de paramètres non-identifiés.

Notez que deux problèmes sont associés à la vérification de l'identification. Le premier est que les estimations limites engendrent elles aussi des déficiences de rang. En d'autres termes, nous ne pouvons pas savoir si une déficience de rang est causée par des limites ou par des paramètres non- identifiés. Les constantes de Bayes empêchent les limites d'apparaître, ce qui résout le premier problème lié à ce message. Cependant, un second problème provient du fait que cette vérification d'identification ne peut pas toujours détecter la non-identification lorsque des constantes de Bayes sont utilisées. En d'autres termes, des constantes de Bayes peuvent faire apparaître en tant qu'identifiés des modèles non-identifiés en réalité.

MESSAGE : nombre maximum d'itérations atteint sans convergence

Cette alerte apparaît lorsque le nombre maximal d'itérations EM et Newton- Raphson est atteint avant de vérifier le critère de tolérance. Si le critère de tolérance (très strict par défaut) est presque atteint, la solution est probablement fiable. Sinon, il est recommandé de ré-estimer le modèle avec une tolérance EM plus sévère et/ou un nombre d'itérations EM plus élevé. Cela retarde le basculement entre itérations EM et itérations de Newton-Raphson. Le nombre d'itérations de Newton-Raphson par défaut (50) est en général suffisant.

MESSAGE : la procédure d'estimation n'a pas convergé (# gradients supérieur à $1.0e-3$)

Ce message peut être lié au message précédent, auquel cas la même solution peut être envisagée. L'absence d'affichage du message précédent suggère un problème plus grave de non-convergence. Les algorithmes pourraient avoir été piégés dans une région très aplatie de l'espace des paramètres (point selle). La meilleure solution est de ré-estimer le modèle avec une graine différente, et si possible avec un plus grand nombre de jeux de valeurs initiales et d'itérations par jeu.

Exemple

Un exemple de régression par les classes latentes est disponible sur le Centre d'aide XLSTAT :

http://www.xlstat.com/demo-lcrf.htm

Bibliographie

Vermunt J.K. (2010). Latent class modeling with covariates: Two improved three-step approaches. Political Analysis, 18, 450-469. Link: http://members.home.nl/jeroenvermunt/lca_three_step.pdf

Vermunt J.K. and Magidson, J. (2005). Latent GOLD 4.0 User's Guide. Belmont, MA: Statistical Innovations Inc. http://www.statisticalinnovations.com/technicalsupport/LGusersguide.pdf

Vermunt J.K. and Magidson, J. (2013a). Technical Guide for Latent GOLD 5.0: Basic, Advanced, and Syntax. Belmont, MA: Statistical Innovations Inc. http://www.statisticalinnovations.com/technicalsupport/LGtechnical.pdf

Vermunt J.K. and Magidson J. (2013b). Latent GOLD 5.0 Upgrade Manual. Belmont, MA: Statistical Innovations Inc.

http://statisticalinnovations.com/technicalsupport/LG5manual.pdf